Universität Leipzig

Fakultät für Mathematik und Informatik
Institut für Informatik

Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung

DIPLOMARBEIT

vorgelegt von
Fabian Schmidt

Leipzig, April 1999

Inhaltsverzeichnis

1. Einleitung

In verschiedenen Bereichen der Linguistik werden Kollokationen genutzt, beispielsweise als Unterstützung bei der Erstellung von Wörterbüchern oder bei der Übersetzung von Fachtexten. Umfangreiche Kollokationssammlungen können aufgrund ihrer Größe nicht manuell erstellt werden. Deshalb wurden in den letzten Jahren verschiedene Verfahren entwickelt, um die Kollokationssammlungen durch statistische Analyse maschinenlesbarer Textkorpora zu erzeugen. Neben guten Kandidaten ermitteln die meisten herkömmlichen Verfahren auch viele Wortpaare, deren Plausibilität nicht unmitttelbar ersichtlich ist. Darum soll in der vorliegenden Diplomarbeit ein neues, in der Abteilung für Automatische Sprachverarbeitung am Institut für Informatik der Universität unter Leitung von Dr. U. Quasthoff entwickeltes Verfahren zur Berechnung von Kollokationen vorgestellt werden.

Aufbauend auf ein Repertoire der nun vorhandenen Kollokationen kann die Fragestellung nach semantischen Relationen zwischen lexikalischen Einheiten weiter ausgebaut werden. Die in dieser Arbeit eingeführten Kollokationen zweiter Ordnung verfolgen diesen Ansatz.

Neben der automatischen Extraktion von Kollokationen stellt auch die adäquate Darstellung derselben ein noch nicht zufriedenstellend gelöstes Problem dar. Für die Visualisierung der im Wortschatzprojekt gefundenen Kollokationen wurde deshalb ein Verfahren entwickelt, das in der Lage ist, eine Menge von Kollokationen in einem Graphen ästhetisch ansprechend und übersichtlich darzustellen.

1.1. Gliederung

Zunächst sollen im Verlauf des ersten Kapitels die verwendeten Fachbegriffe vorgestellt und näher erläutert werden. Im Abschnitt Abschnitt 1.2 wird der Begriff der Kollokation eingeführt. Nachdem die verschiedenen Aspekte des Kollokationsbegriffs in der Linguistik vorgestellt werden, wird auf die Verwendung im weiteren Verlauf der Arbeit eingegangen.

Der darauf folgende Abschnitt 1.3 stellt das Projekt *Deutscher Wortschatz* vor. Das

Projekt beschäftigt sich mit Ausbau und Pflege der zur Zeit wohl umfangreichsten Sammlung von Vollformen deutscher Wörter. Auf der Basis der Wortsammlung werden weitere Anwendungsmöglichkeiten bereitgestellt. Die Datensammlung dieses Projektes lieferte die Grundlagen für die in dieser Arbeit durchgeführten Untersuchungen und Berechnungen zu den Wortrelationen.

Der letzte Abschnitt des ersten Kapitels stellt einige Nutzungsmöglichkeiten für die Kollokationen vor. Es folgen Beispiele für die Verwendung der anderen Relationen, die auf Kollokationen aufbauen.

Das zweite Kapitel beschreibt die theoretischen Hintergründe und die Ursprünge zu linguistischen Konzepten. Zunächst werden die Kollokationen auf Prozesse bei der Sprachverarbeitung im menschlichen Gehirn zurückgeführt und eine klassische Vorgehensweise der Ermittlung von Kollokationen vorgestellt.

Der zweite Abschnitt führt in die Meaning-Text Theory ein, in der die Relationen zwischen lexikalischen Einheiten nach verschiedenen lexikalischen Funktionen klassifiziert werden.

Nachdem herkömmliche Maße der statistischen Bestimmung von Kollokationen vorgestellt worden sind, beschäftigt sich Abschnitt 3.1.2 mit der Herleitung eines neu entwickelten Signifikanzmaßes, das die Grundlage für die im weiteren Verlauf der Arbeit durchgeführten Untersuchungen bildet.

Kollokationen sind auf Wörter beschränkt, die in den Korpora innerhalb einer kleinen Umgebung, wie z. B. im selben Satz, auftauchen. Im Abschnitt 3.2 wird dargestellt, wie die Kollokationen genutzt werden können, um daraus weitere Relationen ableiten zu können. Diese sind nicht mehr auf räumlich benachbarte Wörter beschränkt.

Die Erkennung von linguistischen Konzepten oder mehrteiligen lexikalischen Einheiten anhand der einzelnen Relationen in Wortpaaren stellt ebenfalls ein Problem dar. Deshalb wird in Abschnitt 3.3 ein Verfahren vorgestellt, um aus den Relationen Graphen ableiten zu können, in denen die Konzepte als einzelne Cluster der Graphen repräsentiert werden.

Anschließend gehen wir darauf ein, wie die automatisch ermittelten Relationen durch explizites Wissen qualitativ weiter verfeinert werden können.

Das vierte Kapitel zeigt, wie die Relationen, die aus ihnen abgeleiteten Graphen und die weiteren Daten des Projektes Deutscher Wortschatz dem Nutzer präsentiert werden. Den Kernpunkt bildet dabei ein Verfahren, mit dem die Graphen aus dem Abschnitt 3.3 effizient in eine zweidimensionale Darstellung überführen können.

Den Abschluß der Arbeit bildet eine Zusammenfassung, wobei die erzielten Ergebnisse kritisch bewertet werden und mögliche Erweiterungen hingewiesen wird.

1.2. Kollokationen

Zwischen den lexikalischen Einheiten eines Satzes oder eines Textes bestehen eine Vielzahl semantischer Zusammenhänge, die in ihrer Gesamtheit die Semantik des Satzes oder Textes formen. Eine (automatische) Extraktion dieser Zusammenhänge setzt normalerweise die Kenntnis der Semantik der einzelnen Bestandteile voraus. Diese ist aber insbesondere bei statistischen oder korpuslinguistischen Verfahren nicht oder nur unzureichend bekannt.

Ein „einfacher“ Ansatz baut auf der Annahme auf, daß semantisch in Beziehung stehende Einheiten in verschiedenen Umgebungen vermehrt gemeinsam auftauchen. Dieses Eigenschaft wird als Kollokation bezeichnet.

Der Begriff *Kollokation* geht auf das lateinische `collocatio` zurück, der auf deutsch *Stellung* oder *Anordnung* bedeutet. Hausmann definiert ihn u. a. in [Hm85, S. 118 ff.] als „typische, spezifische und charakteristische Zweierkombinationen von Wörtern“. Bei dem Versuch einer präziseren oder gar allgemeingültigen Definition des Kollokationsbegriffes stößt man schnell auf Schwierigkeiten:

- Der Kollokationsbegriff ist in der Sprachwissenschaft geteilt. Kollokationen werden zum einen auf syntaktisch-semantischer Ebene und zum anderen auf statistischer Ebene untersucht. Kollokationen, die mit Hilfe von statistischen Methoden gefunden werden, umfassen beliebige Wortkombinationen ungeachtet ihrer grammatischen Wohlgeformtheit, während Kollokationen nach dem syntaktischen Ansatz auf die Kombination bestimmter Wortarten (z. B. Substantiv-Adjektiv-Kollokationen) beschränkt sind.

- Kollokationen werden in verschiedenen sprachwissenschaftlichen Schulen wie dem Strukturalismus, der Transformationsgrammatik und dem Kontextualismus untersucht und sind in verschiedenen Bereichen der Linguistik von Bedeutung (z. B. in der Lexikologie und Lexikographie, in den Bereichen der Wortbildung, Fremdsprachendidaktik, Übersetzungswissenschaft, Computerlinguistik und Fachsprachenforschung). In den sprachwissenschaftlichen Teilbereichen haben sich im Laufe der Zeit auf Grund der verschiedenen Sichtweisen unterschiedliche Kollokationskonzepte entwickelt, die jeweils eingene Bezeichnungen prägten.

Wichtige Vertreter der Kollokationsforschung sind u. a. Hausmann [Hm85], Benson, Kohn, Lemnitzer, Greenbaum und Peña. Einen guten Überblick darüber gibt Andrea Lehr in ihrer Dissertation [Lr96], eine kurze Vorstellung der einzelnen Richtungen findet sich auch in Thiessen [Ti99].

Wir gehen vom statistischen Ansatz aus und verstehen unter der *Kollokation eines Wortes* die Wörter, die signifikant häufig mit diesem Wort in einer gewissen

Umgebung — dem Kotext des Wortes — erscheinen. In der Literatur finden sich verschiedene Definitionen der Umgebung eines Wortes, die die Bedeutung des Kollokationsbegriffs nachhaltig beeinflussen. Eine übliche Wahl dieser Umgebung ist das *Wortfenster*, das aus einer festen Anzahl vorausgehender und nachfolgender Wörter im Text besteht.

Im Rahmen des Wortschatzprojektes wurde die Größe der Umgebung wie bei GREENBAUM auf einen Satz festgelegt, da zu den Wörtern Beispielsätze gespeichert werden und in den Sätzen semantisch verwandte Wörter auch weit entfernt voneinander auftreten können. Dies tritt in der deutschen Sprache besonders häufig bei Präfixverben und mehrfach verschachtelten Sätzen auf. Durch die Beschränkung der Umgebung auf einen Satz werden die Kollokationen eingegrenzt. Die Kollokationen im gleichen Satz besitzen eine höhere Aussagekraft als solche, die sich über Satzgrenzen hinaus erstrecken. Wir vermeiden so außerdem Probleme mit nicht signifikanten Kollokationen, die dadurch entstehen, dass das Wortfenster über Absatz- oder Textgrenzen reicht.

Der Abschnitt Abschnitt 3.2 auf Seite 35 beschäftigt sich mit der Aufhebung der Einschränkung der Kollokationen auf Wörter im gleichen Satz, verwendet aber einen anderen Ansatz als die Verwendung eines Wortfensters. Er ist nicht mehr auf räumliche Nähe beschränkt, sondern ermittelt auch Relationen aus verschiedenen Texten.

Die Kollokationen zu einem Wort werden mit Hilfe eines Signifikanzmaßes gefunden; alle Kollokationen die einen bestimmten Schwellenwert überschreiten, heißen *signifikante Kollokationen*. Die Sigifikanzmaße werden im folgenden Abschnitt 3.1 auf Seite 21 vorgestellt.

Die Gesamtheit aller signifikanten Kollokationen zu einem Wort bezeichnen wir als die *Kollokationsmenge* zu diesem Wort, die Gesamtheit aller zugehörigen Signifikanzmaße als *Kollokationsvektor*.

Hinsichtlich ihrer Herkunft lassen sich verschiedene Kollokationstypen unterscheiden. Diese Unterteilung variiert bei den verschiedenen Autoren. Hier wird die Einteilung nach LEMNITZER [Lm97, S. 86] vorgestellt:

- *Komplexe Funktionswörter* sind lexikalische Einheiten, die genau eine grammatische Funktion erfüllen, wie z. B. *sowohl ... als auch* oder *manch ein*.
- Zu *eingliedrigen lexikalischen Zeichen mit einer komplexen Binnenstruktur* zählen z. B. Partikelverben im deutschen (wie *aufhören*), reflexive Verben im spanischen und italienischen und Nomen in skandinavischen Sprachen. Die Teile treten in manchen Fällen im Text getrennt auf.

  Diese Strukturen lassen sich ebenso wie komplexe Funktionswörter leicht im Textkorpus finden, da sie in dieser Funktion immer gemeinsam auftreten, auch wenn sie im Satz verteilt stehen können.

- Bei *mehrgliedrige lexikalische Zeichen*, denen als Ganzes eine Bedeutung zugeschrieben wird, lässt sich diese nicht aus der Bedeutung der Bestandteile rekonstruieren (höchstens mit Kenntnis der Etymologie der Teile). Zu diesen gehören idiomatische Wendungen oder Phraseme (z. B. *die Katze im Sack kaufen, rote Zahlen schreiben*) Die oftmals hohe Binnenvarianz der Phraseme kann die automatische Identifikation erschweren.

- *Mehrgliedrige lexikalische Zeichen mit kompositioneller Bedeutung* sind Kollokationen im syntaktischen Sinn. Sie unterscheiden sich von den freien („unfixierten“) Verbindungen lexikalischer Zeichen durch die Arbitrarität der gegenseitigen Selektion. Ein Kollokator bindet ein oder mehrere lexikalische Zeichen als Kollokanten zuungunsten anderer, bedeutungsgleicher oder -ähnlicher lexikalischer Einheiten.

1.3. Projekt Deutscher Wortschatz

Zu Beginn der neunziger Jahre wurde in der Abteilung für Automatische Sprachverarbeitung am Institut für Informatik der Universität damit begonnen, eine Liste der Wörter der deutschen Sprache aufzubauen, um den zu diesem Zeitpunkt bestehenden Mangel an frei verfügbaren Daten zum deutschen Wortschatz zu beheben. Gesammelt wurden alle Vollformen von Wörtern aus maschinenlesbar verfügbaren Texten zusammen mit ihrer Auftretenshäufigkeit und den mitunter vorhandenen Grammatik- und Sachgebietsangaben, mit dem Ziel, im Laufe der Zeit eine möglichst vollständige Sammlung aller verfügbaren Informationen zu den Wörtern im deutschen Sprachgebrauch aufzubauen.

Ähnliche Zielsetzungen führten zu umfangreichen manuell erstellten Sammlungen, die nur mit hohem Aufwand an den sich ständig ändernden Sprachgebrauch angepasst werden können. Als Beispiel hierfür sei Dornseiffs Buch *Der Deutsche Wortschatz nach Sachgruppen* erwähnt, dessen Anspruch der Autor in [Do64, S. 41] so formulierte:

> Für die Einzelbegriffe sollte nun möglichst alles aufgeführt werden: Gottseliges, Schnodderiges, Fremdwörter, Papierenes, Menschlich-Allzumenschliches, Derbes, was Snobs sagen, die Backfische, Soldaten, Schüler, Kunden (Rotwelsch), Seeleute, Studenten, Gelehrte, Jäger, Börsianer, Pfarrer, die Zeitungen, wie sich der Gebildete ausdrückt im täglichen Verkehr, im Honoratiorendeutsch, in der gehobenen Literatursprache.

Mit dem Anwachsen der Datensammlung fanden sich unter den Wörtern vermehrt solche, bei denen nicht festgestellt werden konnte, ob es sich um fehlerhaft geschrie-

bene Wörter, zulässige Varianten, Eigennamen oder Fachbegriffe handelt. Deshalb wurde seit Anfang 1996 für jede Form ein Belegsatz gesammelt.

Diese Informationen ermöglichen neben Untersuchungen zur morphologischen Zerlegung von Wörtern, der automatischen Ergänzung von Grammatikangaben und Auffindung von korrekten Schreibweisen auch die Analyse von Kollokationen anhand eines großen Korpus. Da zunächst nur Sätze gesammelt wurden, in denen ein neues Wort auftauchte, erwies sich die Sammlung als nicht repräsentativ und daher für Berechnung von Kollokationen wenig geeignet. Deshalb sammelten wir seit 1998 alle Sätze, bis zu einer bestimmten Länge, die eindeutig aus dem Text separiert werden konnten (ignoriert wurden etwa Überschriften oder Einleitungen vor direkter Rede). Ein Volltextindex ermöglicht eine effiziente Suche in den Sätzen.

Die Kollokationsberechnungen sind stark abhängig von der zu Grunde liegenden Textbasis. Die einzige Beschränkung des breit gefächerten allgemeinsprachlichen Korpus besteht darin, dass die einzelnen Texte einer bestimmten Zeitepoche entstammen und damit den damaligen Wissensstand und Sprachgebrauch repräsentieren. Bei der Suche nach Kollokationen von Fachbegriffen erzielt man nur für die Bereiche gute Resultate, für die bereits fachsprachliche Texte eingelesenen wurden. In Zukunft wird die Textbasis durch Fachtexte aus anderen Bereichen wie z. B. der Medizin und Physik weiter verbessert werden.

Die folgende Übersicht stellt eine Auswahl der in das Wortschatz-Lexikon eingearbeiteten Korpora dar:

- allgemeinsprachliche Texte
  Donau-Kurier 1992-1993
  Frankfurter Allgemeine Zeitung 1994
  Frankfurter Rundschau 1992
  Süddeutsche Zeitung 1995-1996
  die tageszeitung 1986-1997
  Die Zeit 1995-1996

- fachsprachliche Texte
  bild der wissenschaft 1993-1996
  Computerzeitung 1993-1996
  Arbeitsrechtliche Praxis
  Neue Juristische Wochenschrift
  Rechtstexte der UB Media
  Deutsche Zeitschrift für Philosophie 1995
  Geographische Rundschau

- Korpora der Sprachwissenschaft
  Bonner Zeitungskorpus
  Limas-Korpus

Mannheimer Korpus 1 und 2
Mannheimer Morgen

- Wörterbücher
- Lexika
  Lexikon des internationalen Films u. a.

1.4. Motivation

Viele Bereiche der Linguisitik beschäftigen sich intensiv mit Kollokationen. Im folgenden sollen einige Anwendungsgebiete für Kollokationen in den einzelnen Bereichen aufgezeigt werden.

Computerlinguistik

Kollokationen stellen mit Beziehungen zwischen lexikalischen Einheiten auch Beziehungen zwischen *Bedeutungen* her. Dadurch kommen sie für die Wissensimplementierung auf dem Rechner in Betracht. Desweiteren sind Kollokationen geeignet, semantisch-paradigmatische Wortfelder zu bilden, die in Frames eingesetzt werden können. Anwendungsbereiche sind damit die maschinelle Sprachverarbeitung, Frage/Antwort- und Expertensysteme.

Die Frage nach der richtigen Kollokation kann man mit der Situation eines nicht muttersprachlichen Übersetzers vergleichen, der vom Muttersprachler auf geringfügig vom allgemeinen Sprachempfinden abweichende Wendungen aufmerksam gemacht wird, der trotz langer Erfahrung und dem Wissen um Synonyme unübliche Wendungen gebraucht. Solche Synonyme sind z. B. *ablehnen, abschlagen, abweisen, verweigern, ...* , die verschiedene Nuancen in der Verwendung aufweisen, aber z. B. im Tschechischen nur ein Äquivalent haben.

In dem Sinne ist die Wortschatz-Wortliste für einen Übersetzer ein „riesengroßes Wörterbuch“ (Kolečková), das alle vorkommenden Kollokationen speichert.

Lexikologie und Lexikographie

Eine Aufgabe dieser Gebiete ist die Beschreibung des Wortschatzes einer Sprache durch Nachschlagewerke. Dabei besteht eine Teilaufgabe in der lernorientierten Darstellung von Kollokationen, die grundlegende Elemente des Wortschatzes sind. Dazu werden ein- und zweisprachige Lernerwörterbücher, Textproduktions-, Rezeptions-,

Kombinations- und Kollokationswörterbücher erstellt, die Kollokationen anführen und erläutern und bei der Suche nach der korrekten Kollokation helfen sollen.

Obwohl Kollokationen bei der Wörterbucharbeit von Fremdspachenlernern die drittwichtigste Problemgröße bilden, sind sie oft nur unsystematisch oder unvollständig erfasst (Kolečková 1997). Mit einer automatischen Bestimmung der Kollokationen aus großen Korpora können diese Angaben ergänzt werden.

Fremdsprachdidaktik

Kollokationen sind wichtige Größen der Fremdsprachendidaktik, da sie als vorgefertigte Wortverbindungen bestimmte Inhalte und Zusammenhänge wiedergeben und durch sie Vergleiche zwischen den Wortverbindungen verschiedener Sprachen möglich sind. Durch Kollokationen wird im ersten Fall ein höheres fremdsprachliches Niveau erreicht und im zweiten der Fremdsprachenlernprozeß unterstützt.

Stabile Wortverbindungen sind in der Regel ein wesentlicher und für den Lerner unumgänglicher Bestandteil einer Fremdsprache. Als Mittel differenzierter Ausdrucksweise auf höherem fremdsprachlichem Niveau sind sie auf folgenden Ebenen mit diesen Formen förderlich:

- sprachliche Kompetenz: Redewendungen
- kommunikative Kompetenz: Gruß- und Höflichkeitsformeln und weitere Wendungen der Kommunikation
- kulturelle Kompetenz: Sprichwörter, Slogans, literarische Anspielungen und Zitate. Diese Form ist auf Grund raum-zeitlicher und mentaler Distanz zum zielsprachigen Land mitunter schwer zugänglich und stellt hohe Anforderungen an den Lernenden.

Kollokationen treten vor allem im Bereich der sprachlichen Kompetenz auf. Trotz ihrer offensichtlichen Bedeutung werden sie jedoch ebenso wie die anderen fest vorgegebenen Formen im Fremdsprachenunterricht marginalisiert, da sie sich im Gegensatz zu den freien Kombinationen nicht hinreichend erklären lassen. Hinzu kommt, dass neben negativem Transfer (Übertragung ausgangssprachlicher Grammatik und Lexik in die Zielsprache), Übergeneralisierung (Simplifizierung fremdsprachlicher Strukturen) und fehlender Vertrautheit mit der Fremdsprache zielsprachliche Sprachelemente vermischt werden. Das Ergebnis sind grammatisch korrekte, doch für den Muttersprachler ungewohnte Ausdrücke und Sätze.

Übersetzungswissenschaften

Kornelius umreißt die Problematik der Kollokation in der Übersetzungswissenschaft, indem er Kollokationen „maligne Mikroeinheiten“ nennt, die sich einfach in die eigene Sprache herübersetzen lassen, sich bei der Übersetzung in die Fremdsprache jedoch zu „Problemgrößen der Produktion“ entwickeln. Belegt wird dies durch Ausdrücke wie *Träume lösen* im Hebräischen oder *Tabletten oder Zigaretten trinken* im Japanischen. Bei der Hinübersetzung der Kollokation wird von der äquivalenten zielsprachlichen Basis ausgegangen, der mehrere mögliche Kollokationen gegenüberstehen, die jedoch nicht alle für die gesuchte Kollokation geeignet sind. Die geeignete Kollokation wird durch semantischen Abgleich ermittelt.

Eine Besonderheit bei Kollokationen ist die Entstehung von Fehlern bei der einzelwörtlichen Übersetzung. Dies wird auf den teilidomatischen Charakter einiger Kollokationen zurückgeführt, dem durch die einzelwörtliche Übertragung nicht Rechnung getragen werden kann. Kollokationen müssen daher als komplexe sprachliche Einheiten betrachtet werden. Dabei sind die Basen ohne Probleme in die Zielsprache übertragbar, die Übersetzung der Kollokationen hingegen muss kontextuelle Bedingungen, wie z. B. die Kommunikationssituation, berücksichtigen.

Fachsprachenforschung

Die Aufgabe der Fachsprachenforschung besteht zum einen in der Generierung von Fachwörtern, die noch nicht versprachlichte Sachverhalte oder noch nicht konventionalisierte Ausdrücke festlegen. Daneben gehen die Ergebnisse der Fachsprachenforschung in Fachwörterbücher ein, die den Wortschatz eines Fachs enthalten. An dieser Stelle überschneidet sich die Fachsprachenforschung mit der Lexikographie und heißt Fachlexikographie.

Kollokationen werden in Fachwörterbüchern jedoch noch zu wenig berücksichtigt. Fachsprachen sind hinsichtlich ihrer grammatischen und lexiaklischen Eigenschaften oft nicht umfassend erforscht, zudem unterliegen sie oft fremdsprachlichem Einfluß.

Bergenholtz & Tarp gehen davon aus, dass Angaben zu Kollokationen in Fachwörterbüchern erstens unabdingbar sind (für die Hinübersetzung) und zweitens die Bedeutung fremdsprachliche Wortverbindungen näherbringen (bei der Herübersetzung). Eine auf die Fachlexikographie anwendbare Kollokationstheorie existiert jedoch noch nicht.

Erschwerend kommt hinzu, daß die in Fachtexten häufigen Mehrwortverbindungen (Verbindungen aus mehreren Wörtern im Gegensatz zu den Komposita) Mehrworttermini oder Kollokationen sein können. Ein Mehrworttterminus ist ein vollständiger Fachausdruck, eine Kollokation dagegen die Verbindung aus einem (ein- oder mehrteiligen) Fachwort und einem oder mehreren Lexemen, die nicht Teil dieses Fach-

worts sind. Mangelnde Fachkenntnis des Übersetzers, der oft nicht zwischen beiden Formen unterscheiden kann, führt deshalb zu Übersetzungsfehlern; der übersetzte Ausdruck ist entweder falsch oder unüblich.

Weitere Relationen

Neben Kollokationen beschäftigt sich die Arbeit mit der Ermittlung von Wörtern aus anderen Realtionen, die sich aus den Kollokationen ableiten lassen. Auch für diese sollen einige Anwendungsmöglichkeiten aufgeführt werden. Für diese Relationen erschließen sich zahlreiche Nutzungsmöglichkeiten; z. B. im *Information Retrieval* und der *Computerlexikographie*:

- **Synonyme:**
  Dem Benutzer eines Retrieval-Systems brauchen nicht alle Synonyme bekannt zu sein bzw. sie müssen nicht alle angegeben werden; die Suche kann automatisch auf sinnverwandte Wörter ausgedehnt werden.
- **Identifizierung komplexer lexikalischer Einheiten:**
  Unter den komplexen lexikalischen Einheiten unterscheiden wir: komplexe Funktionswörter (sie bestehen aus mehreren einzelnen Wörtern, wie z. B. *unter anderem*, die genau eine grammatikische Funktion erfüllen), Mehrwortbegriffe[1] (z. B. *Brandenburger Tor*), Partikelverben (zusammengesetzte Verben, die im Satz in mehrere Bestandteile zerfallen) und Phraseme und idiomatische Wendungen. Zu Wörtern aus diesen komplexen lexikalischen Einheiten können automatisch die anderen Mitglieder der Einheit erkannt werden.
- **Thematische Klassifizierung:**
  Ausgehend von bereits klassifizierten Fachwörtern können andere Wörter, die signifikant häufig mit diesen in Texten auftauchen, dem gleichen Sachgebiet zugeordnet werden. Darauf aufbauend können Texte klassifiziert werden, etwa eingehende Ticker-Meldungen in einer Zeitungsredaktion.
- **Vervollständigung semantischer Cluster:**
  Zum Aufbau eines onomasiologischen Wörterbuchs können aus den einzelnen Relationen unter Vorgabe typischer Kandidaten semantische Cluster vervollständigt werden.
- **Spreading-Activation-Netze:**
  Zur Verwendung in einem Retrieval-System kann aus den einzelnen Relationen ein Netz von Termassoziationen generiert werden. Es dient u. a. dazu,

[1] Einige Autoren schränken Kollokationen auf Mehrwortbegriffe ein. Siehe dazu die Diskussion des Kollokationsbegriff im Abschnitt 1.2 auf Seite 5

Dokumente zu bestimmten Termen zu finden. Verschiedene Ansätze werden in [Ru95, S. 181 ff.] vorgestellt.

- **Allgemeinlexikon:**
  Die Relationen stellen auch einen Wissensspeicher dar, der z. B. über typische Merkmale (Vorname, Beruf, Wirkungsbereich etc.) prominenter Personen oder räumlich benachbarte geographische Einheiten (Städte, Hauptstädte, Länder etc.) Auskunft gibt.

- **Head-Modifier-Strukturen:**
  Gefundene Head-Modifier-Strukturen können verwendet werden, um ein *Combinatory Dictionary* aufzubauen, wie es Benson, Benson & Illson für das Englische aufgebaut haben. Der Aufbau beruht auf den lexikalischen Funktionen von Mel'čuk, siehe dazu Abschnitt 2.2 auf Seite 19.

2. Theorien zu linguistischen Konzepten

Schon bevor Text automatisch verarbeitet werden konnte, beschäftigten sich Linguisten mit Kollokationen. In diesem Kapitel wird zunächst der psychologische Hintergrund der Kollokationen – die Assoziationen – vorgestellt und gezeigt, wie diese automatisch bestimmt werden können.

Im Anschluss daran stellen wir eine moderne Theorie zu linguistischen Konzepten vor und diskutieren deren Ergebnisse und Einsatzmöglichkeiten.

2.1. Klassische Vorgehensweise zur Ermittlung von Assoziationen in der Psychologie

Beim Aufbau eines onomasiologischen Lexikons ist man auf das Fachwissen profesioneller Lexikographen angewiesen, die auf den zu behandelnden Gebieten über eine hohe Sachkompetenz verfügen müssen. Analog benötigen Profi-Rechercheure für die effiziente Suche in großen Dokumentenkollektionen eine besonders ausgeprägte Kompetenz bei der Schlagwortassoziation. In solchen Anwendungsfällen soll die automatische Auffindung von Relationen helfen. Das Modell von Wettler und Rapp [Rp96] zur automatischen Assoziation von zusätzlichen Schlagwörtern beruht auf der Theorie des *klassischen Assoziationismus*, in die an dieser Stelle einen Einblick gegeben werden soll.

Die assoziative Arbeitsweise des menschlichen Gedächtnisses wurde früh erkannt. In dem von Galton (1880) eingeführten Assoziationsexperiment wurde erstmals versucht, das Assoziationsverhalten von Menschen systematisch zu erfassen. Hierzu mußten Versuchspersonen auf ein einzelnes vorgegebenes Wort, den Stimulus, mit dem anderen Wort antworten, das ihnen zuerst einfiel. Auf diese Weise ergaben sich Tabellen der Häufigkeiten, mit denen verschiedene assoziative Antworten auf bestimmte vorgegebene Stimuluswörter gegeben wurden. Solche Tabellen, wie sie später beispielsweise von Russell & Jenkins (Jenkins, 1970) erfaßt wurden,

werden als Assoziationsnormen bezeichnet.

Zur Erklärung des in diesen Assoziationsnormen dokumentierten Verhaltens werden in der Literatur eine Vielzahl unterschiedlicher Mechanismen angenommen, die der Speicherung im Gedächtnis zu Grunde liegen sollen (vergleiche die Klassifizierung der Assoziationen nach Jung & Ricklin in Tabelle 2.1).

Die Assoziationsklassen der unteren vier Gruppen sind bei einer automatischen Assoziation weder erwünscht noch reproduzierbar, da sie nicht auf zeitlicher Kontiguität basieren. Die Erzeugung von Klangreaktionen ist zwar abhängig von der verwendeten Textbasis (ob in ihr etwa Gedichte oder Liedverse enthalten sind), jedoch sind automatisch reproduzierte Klangassoziationen in der Regel auf andere Assoziationsklassen wie Sprichwörter rückführbar.

Die Grenze der Reproduzierbarkeit durch andere Versuchspersonen liegt bei persönlichen oder episodischen Relationen. Solche Produktionen sind für andere Personen nicht nachvollziehbar, da sie auf einem gemeinsamen Vorkommen in einer persönlichen Episode fußen.

Der Physiologe David Hartley (1749) vertrat bereits Mitte des 18. Jahrhunderts die Ansicht, daß sich eine Vielzahl vermuteter Assoziationsgesetze auf ein einziges reduzieren ließen, nämlich auf das Assoziationsgesetz durch zeitliche Kontiguität: Ähnliche Objekte werden häufig gleichzeitig oder in unmittelbarer Folge wahrgenommen. Sehr klar wurde das Kontiguitätsprinzip von William James im Jahre 1890 formuliert:

> Objects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before. This statement we may name the law of mental association by contiguity.
> *(James, 1890, S. 561.)*

Den Kern dieser Aussage findet man auch bei Ebbinghaus (1919, S. 678):

> ... wenn beliebige seelische Gebilde einmal gleichzeitig oder in naher Aufeinanderfolge das Bewußtsein erfüllt haben, so ruft hinterher die Wiederkehr einiger Glieder des früheren Erlebnisses Vorstellungen auch der übrigen Glieder hervor, ohne daß für sie die ursprünglichen Ursachen gegeben zu sein brauchen.

In der heutigen Psychologie wird überwiegend die Ansicht vertreten, daß das Kontiguitätsgesetz nicht ausreiche, um die im Assoziationsversuch ermittelten Wortassoziationen zu erklären. Wettler (1980, S. 34) interpretiert experimentelle Ergebnisse

Assoziationsklasse	Häufigkeit	Beispiel
Innere Assoziationen		
Koordination zwischen Reiz und Antwort	19,6 %	
Beiordnung		Kirsche – Apfel
Unterordnung		Baum – Buche
Überordnung		Katze –Tier
Kontrast		süß – sauer
Prädikative Beziehung	18,7 %	
Substantiv und Adjektiv		Schlange – giftig
Substantiv und Verb		Harz – klebt
Bestimmung von Ort, Zeit, Mittel und Zweck		essen – Mittag
Definitionen oder Erklärungen		Türe – Hauptwort
Kausale Abhängigkeit	1,0 %	Schmerz – Tränen
Äußere Assoziationen		
Koexistenz	16,0 %	Schüler – Lehrer
Identität	6,3 %	großartig – prächtig
Sprachlich-motorische Formen	26,5 %	
eingeübte sprachliche Verbindung		dunkel – hell
Sprichwörter und Zitate		Glück – Glas
Wortzusammensetzungen und -veränderungen		Tisch – Bein
vorzeitige Reaktion (die Antwort bezieht sich lediglich auf den ersten Teil des Reizwortes)		dunkelrot – hell
Interjektionen		stinken – pfui
Klangreaktionen		
Wortergänzung	1,1 %	Wunder – bar
Klang	2,2 %	rosten – Roastbeef
Reim	0,8 %	Herz – Schmerz
Restgruppe		
Mittelbare Assoziationen (die Beziehung zwischen Reiz und Antwort ist durch ein drittes Wort vermittelt)	1,2 %	weiß – weit
sinnlose Reaktion	0,3 %	
fehlende Reaktion	1,5 %	
Wiederholung des Reizwortes	0,1 %	
Egozentrische Reaktionen	1,7 %	tanzen – mag ich nicht
Perseveration (die Antwort steht in Beziehung zu einem früher gegebenen Reizwort)	1,2 %	Ratte – Korb
Wiederholung einer früher gegebenen Antwort	9,1 %	

Tabelle 2.1.: Klassifizierung der Assoziationen nach Jung & Ricklin (1906)

mit sinnlosen Silben von Foppa (1963) wie folgt: „Daß die zeitliche Aufeinanderfolge den einzigen Faktor bilde, durch welchen die Verknüpfung von Elementen im Gedächtnis bestimmt wird, gilt inzwischen als widerlegt.“ Matthäus (1980, S. 624) kommt zum Ergebnis, dass die im Assoziationsexperiment gefundenen Beziehungen zwischen Wörtern außerhalb dieser experimentellen Situation nicht beobachtbar seien, und dass der Versuch deshalb kein geeignetes Instrument für die Untersuchung sprachlicher Prozesse sei. Assoziationen seien deshalb „ ... als Phänomen uninteressant und als Modelle für anderes Verhalten ungeeignet.“ Jenkins (1974) kommt in seinem Aufsatz „Remember that old theory of memory? Well, forget it!“ zu der Auffassung, daß die Assoziationstheorie keine brauchbaren Ergebnisse geliefert hätte. Nach Clark (1970) sind freie Assoziationen das Ergebnis von symbolischen informationsverarbeitenden Prozessen. Dabei werde das Stimuluswort zunächst semantisch kodiert und darauf durch semantische Transformationen die assoziative Antwort abgeleitet.

Demgegenüber konnte Rapp in [Rp96] nachweisen, dass sich die bei Versuchspersonen gefundenen freien Wortassoziationen allein auf der Grundlage des Assoziationsgesetzes in guter Näherung vorhersagen lassen. Ausgangspunkt sind zwei Annahmen, die sich aus dem Assoziationsgesetz ableiten lassen, wenn dieses auf einzelne Wörter bezogen wird (vergl. Rapp & Wettler, 1992b):

- Beim Erlernen einer Sprache werden zwischen denjenigen Wörtern hohe Assoziationsstärken aufgebaut, die in rezipierter Sprache häufig in dichter zeitlicher Aufeinanderfolge auftreten.
- Die so gelernten Assoziationen bestimmen den thematischen Ablauf beim Generieren von Sprache: Es können nur solche Inhaltswörter in dichter zeitlicher Aufeinanderfolge ausgesprochen bzw. niedergeschrieben werden, die untereinander (oder mit externen Stimuli) hohe assoziative Verbindungsstärken aufweisen.

In Tabelle 2.2 auf der nächsten Seite werden die Antworten aus dem Assoziationsexperiment von Russell & Meseck (1959) mit den Kollokationen aus dem Wortschatzprojekt und den Vorhersagen von Rapp verglichen. Zu ausgewählten Stimuli werden die zehn häufigsten Antworten angeführt. Der Zahlenwert gibt an, wie viele der 60 Versuchspersonen diese assoziative Antwort genannt hatten. In die Berechnung der Kollokationswerte aus dem Wortschatzprojekt wird in Kapitel Abschnitt 3.1.2 auf Seite 25 eingeführt. Rapp stellt sein Berechnungsverfahren in seiner Dissertation [Rp96] vor.

An den Assoziationen zu *Mond* kann man deutlich den zeitlichen Abstand des Assoziationsexperimentes und der Textbasen der maschinellen Berechnungen erkennbar. So assoziiert ein Drittel der Versuchspersonen *Sputnik*, wohingegen in den Neunzigern die Marsexpedition stärker im Vordergrund stehen. Die in Russland vieldisku-

Stimulus	Antwort	Anzahl VPn.	Kollokation	Wert	Antwort nach Rapp	Assoz.-Stärke
Butter	Brot	60	Brot	51	Brot	2,88
	weich	40	Käse	49	Eier	2,50
	Milch	32	Zucker	29	Gramm	1,74
	Margarine	27	Milch	23	Milch	1,71
	Käse	20	Margarine	22	Margarine	1,54
	Fett(e)	16	Mehl	18	Zucker	1,23
	gelb	14	Eier	16	Obst	1,10
	Butterbrot	8	Pfund	14	Küche	0,88
	Dose	6	zerlassener	13	Geruch	0,87
	essen	6	Fleisch	13	Fisch	0,84
Mond	Stern(e)	46	Sonne	81	Mars	0,62
	Sonne	39	Erde	49	Landung	0,57
	Nacht	30	Planeten	23	Sonne	0,49
	Sputnik	19	Omon	19	Erde	0,47
	Schein	17	hinterm	17	Rakete	0,31
	Sichel	13	Sterne	15	landen	0,29
	rund	11	Himmel	14	Astronauten	0,27
	Rakete	9	Silberner	14	Planeten	0,22
	Gestirn	8	Mars	11	Sterne	0,21
	Erde	6	Damenchor	9	Sonde	0,21
rot	grün	38	blau	87	gelb	2,32
	Farbe	24	grün	75	blau	1,89
	blau	22	gelb	55	grün	1,80
	gelb	19	schwarz	42	schwarz	1,03
	weiß	17	gefärbt	19	grüne	0,94
	schwarz	16	weiß	14	Farben	0,53
	Liebe	14	färbt	13	grünen	0,51
	Stier	10	orange	13	Gold	0,48
	grell	9	leuchtet	13	Koalition	0,48
	Blut	7	braun	13	Hessen	0,47
schlafen	Bett	57	schlafe	68	nachts	3,56
	wachen	45	ruhig	47	wachen	3,12
	gehen	18	nachts	44	ruhig	1,61
	Ruhe	18	essen	37	Bett	1,53
	müde	17	einschlafen	29	Nacht	1,51
	träumen	14	schlaft	27	müde	1,47
	Nacht	12	schläfst	25	tagsüber	1,42
	ruhen	10	durchschlafen	22	Katze	1,30
	essen	6	Nacht	18	Schlaf	1,18
	schnarchen	6	ausschlafen	18	wach	1,13

Tabelle 2.2.: Vergleich Assoziationsexperiment – Wortschatz – Rapp

tierte Satire *Omon hinterm Mond* taucht erst 1998 häufig in deutschen Zeitungstexten auf. Da das Wort *Omon* jedoch selten ist, aber in 11 von 22 Fällen zusammen mit *Mond* im Text steht, ist diese Verbindung signifikant. Analog erklärt sich die Assoziation zum *Haidhauser Damenchor „Silberner Mond“*. Hingegen setzt Rapp voraus, dass die Versuchspersonen in der Regel mit geläufigen Wörtern antworten und unterdrückt deshalb Wörter mit niedriger Korpushäufigkeit.

Die Bevorzugung von *Koalition* und *Hessen* vor anderen Farben unter den Kollokationen von *rot* führt Rapp auf Korpuseinflüsse zurück: die Zeitungstexte sind zu einem hohen Teil politisch geprägt. In den im Wortschatz-Projekt verwendeten Kollokationen tauchen sie nicht auf, da wir zur Kollokationsberechnung keine Stammformreduktion durchführen.

Weiterhin setzt Rapp durch die Verwendung eines asymmetrischen Bewertungsmaßes einen stärkeren Bezug zu intellektuell bestimmten Assoziationen.

2.2. Meaning-Text Theory

Die Meaning-Text Theory, die von Igor A. Mel'čuk in [Ml76] eingeführt wurde, beschäftigt sich hingegen nicht mit untypisierten Relationen, sondern kategorisiert die Relationen zwischen Wörtern in *lexikalischen Funktionen* und ist bemüht, zu den einzelnen Wörtern möglichst vollständig die lexikalischen Einheiten zusammenzustellen, die in einer solchen lexikalischen Relation stehen.

Aus den lexikalischen Einheiten soll ein sprachunabhängiges Modell lexikalischer Funktionen aufgebaut werden. Zwischen einzelnen Sprachen gibt es aber keine 1:1-Entsprechung einzelner Wörter. Die Meaning-Text Theory betrachtet deshalb *lexikalische Einheiten*, wie z. B. Lexeme oder Phraseme. Diese Einheiten lassen sich wiederum hinsichtlich ihrer sprachlichen Verwendung in zwei Gruppen aufteilen.

Die erste, größere Gruppe bilden die *semantisch basierten* lexikalischen Einheiten. Bei der Textproduktion werden sie auf Grund ihrer Bedeutung ausgewählt, unabhängig von den anderen Einheiten des Textes. Beispielsweise zählt dazu ein *Raubtier mit langschnäuzigem Schädel, buschigem Schwanz, nacktem, feuchtem Nasenspiegel und nicht zurückziehbaren, stumpfen Krallen – der Hund*. Wenn im Text die Sprache auf Hunde kommt, wird ein Sprecher auf Grund seines Weltwissens dieses Wort wählen.

Andere lexikalische Einheiten wählt ein Sprecher in Abhängigkeit von Einheiten, über die er gerade spricht. Er greift nicht über die Bedeutung der Einheiten auf diese zu, sondern durch Beziehungen von anderen, schon ausgewählten Einheiten. Diese Beziehungen sind in seinem lexikalischen Gedächtnis gespeichert.

Die Auswahl dieser Einheiten ist also *lexikalisch basiert*. Sie erfolgt dabei entlang

der *paradigmatischen* oder der *syntagmatischen* Achse:

Zum einen will der Sprecher vielleicht ein Draht- oder Ledergeflecht erwähnen, dass vor dem Maul des Hundes befestigt wird und so ein zubeißen desselben verhindert – der Beißkorb; oder ein als Hundefutter industriell hergestelltes, vitaminreiches hartes Gebäck – der Hundekuchen. Beides sind Beispiele für paradigmatisch gewählte Lexeme, die im lexikalischen Speicher des Sprechers nicht über ihre semantische Bedeutung, sondern über eine Relation z. B. zu Hund referenziert werden.

Andererseits möchte der Sprecher Ereignisse erwähnen, die mit Hunden in Beziehung stehen: wenn der Hund tiefe Warnlaute von sich gibt, dann knurrt er. Diese Wahl ist syntagmatisch bedingt (beruht also auf einer Verbindung zweier lexikalischer Einheiten zu einer größeren Einheit).

Die paradigmatischen und syntagmatischen Funktionen nach Igor A. Mel'čuk [Ml76] werden im Anhang A auf Seite 74 einzeln aufgeführt.

Ein Hauptanwendungsgebiet der Meaning-Text Theory ist der Aufbau beschreibender Lexika mit Angabe der möglichen Verknüpfungen der Lexeme (Explanatory-Combinatorial Dictionaries) wie dem *Combinatory Dictionary of English* von Benson, Benson & Ilson. In solchen Wörterbüchern sind die Kombinationsmöglichkeiten von Lexemen aufgeschlüsselt. Wie bereits in Abschnitt 1.4 erwähnt wurde, können diese z. B. zum Erlernen der Sprache oder zur Übersetzung eingesetzt werden.

Darüber hinaus werden die lexiko-syntaktischen Ideen der Meaning-Text Theory genutzt, um beim Parsen von Text Mehrdeutigkeiten aufzulösen (siehe dazu die Arbeiten von Alexander Nakhimovsky), oder um die Repräsentation von Lexemen in verschiedenen Sprachen zu vergleichen.

Es ist bis jetzt nicht möglich, alle automatisch ermittelten Relationen lexikalischen Funktionen zuzuordnen. Jedoch ist es für den Lexikographen eine Hilfe, dass die potentiellen Kandidaten zum Aufbau des Lexikon-Eintrags zur Auswahl dargeboten werden können.

3. Korpuslinguistische Ansätze zur Erkennung und Verknüpfung linguistischer Konzepte

Die Erkennung der vielfältigen Zusammenhänge von linguistischen Konzepten auf den verschiedenen Ebenen des Sprachsystems stellt seit jeher eines der Hauptziele der Linguistik dar.

Kollokationen lassen sich als prominente Vertreter dieser Klasse nur durch systematische Analyse möglichst vieler Texte sicher identifizieren. Da bei einer solchen Analyse einer manuellen Vorgehensweise allein durch den Umfang der Dokumentensammlung Grenzen gesetzt sind, kommen hier immer häufiger automatische Verfahren zum Einsatz.

In diesem Kapitel stellen wir einige der bekanntesten Ansätze vor und vergleichen deren Leistungsfähigkeit. Darüber hinaus werden wir ein eigenes Verfahren zur Auffindung von signifikanten Kollokationen herleiten und dessen Eignung für das Wortschatz-Projekt rechtfertigen.

3.1. Ermittlung signifikanter Kollokationen

Zur Berechnung von Kollokationen wird zunächst ein Kollokationsmaß benötigt, um die einzelnen lexikalischen Einheiten zueinander in Beziehung setzen zu können. Überschreitet dieses Maß einen gewissen Schwellwert, so gehen wir davon aus, dass es sich um eine signifikante Kollokationen handelt.

Um ein geeignetes Kollokationsmaß zu finden, haben wir innerhalb des Wortschatz-Projekts untersucht, wie verschiedene, intellektuell ausgewählte Wortpaare, die zueinander signifikante Kollokationen sind, von den verschiedenen Maßen bewertet werden. Dafür wurde eine Menge von 7000 Kollokationspaaren gebildet, für die die entsprechenden Werte einiger vielversprechender Maße ermittelt wurden.

3.1.1. Überblick über herkömmliche Maße

Aus den in der Literatur bekannten Maße wurden die ausgewählt, die interessante Ergebnisse versprachen und von denen der exakte Algorithmus zur Berechnung bekannt war.

Zunächst sollen die in diesem Abschnitt verwendeten Bezeichungen eingeführt werden:

- Wörter: a, b, i
- Anzahl aller Wörter in den betrachteten Korpora: n
- Auftretenshäufigkeit des Wortes a im Korpus: $H(a)$
- Auftretenshäufigkeit der Kollokationen a, b im Korpus: $H(a, b)$ (Anzahl der Sätze, in denen beide Wörter auftreten)
- Auftretenswahrscheinlichkeit eines Wortes: $P(a) = H(a)/n$
- Wahrscheinlichkeit, dass Wörter a und b gemeinsam im Wortfenster bzw. Satz stehen (Wahrscheinlichkeit des *Kovorkommens*): $P(a, b) = H(a, b)/n$

Mutual Information Index

Der Mutual Information Index wurde von Fano entwickelt und von Church et al. (1991) in seiner jetzigen Fassung formuliert.

Der Wert des *Mutual Information Index* $\sigma_{MI}(a, b)$ ist bestimmt durch die Differenz aus der Wahrscheinlichkeit des Kovorkommens zweier Wörter $P(a, b)$ und der Wahrscheinlichkeit des Vorkommens der beiden Wörter unabhängig voneinander. Zur effizienten Berechnung wird die Formel so umgestellt, dass nur einmal logarithmiert werden muss.

$$\sigma_{MI}(a, b) = \log_2 P(a, b) - \log_2(P(a)P(b)) = \log_2 \frac{P(a, b)}{P(a)P(b)} \tag{3.1}$$

Wenn a und b statistisch unabhängig sind, ist der Quotient der beiden Ausdrücke 1. Durch die Logarithmierung ergibt sich eine klare Trennung an der Stelle Null. Ist der Wert des Mutual-Information-Indizes größer als Null, so sind a und b voneinander abhängig in dem Sinn, dass sie häufiger als zufällig zusammen auftreten. Umgekehrt ist σ_{MI} kleiner Null, wenn a und b seltener als zufällig zusammen auftreten.

Durch den Mutual Information Index werden Kollokationen seltener Wörter stark überbewertet. Dies steht in keinem Zusammenhang zu der Annahme von RAPP in [Rp96] für die Bildung von Assoziationen beim Menschen, nach der häufige Wörter stärkere Assoziationen bilden. Außerdem wird für seltene Wörter, die zufällig gemeinsam in einem Satz auftauchen, eine hohe Bewertung getroffen, die sich bei einem größeren Datenbestand nicht halten lässt. So ergibt sich z. B. für zwei Wörter, die nur einmal im Text auftauchen, dann aber im selben Satz stehen, bei einer Korpusgröße von n Wörtern der folgende Indexwert:

$$\begin{aligned} \sigma_{MI} &= log_2 \frac{1/n}{1/n \cdot 1/n} \\ &= log_2 n \end{aligned}$$

Für Wörter, die je 1000 Mal im Korpus stehen und jedes Mal gemeinsam im Satz auftauchen, damit aber gegenüber dem obigen Beispiel erheblich signifikanter sind, ergibt sich aber nur ein kleinerer Wert:

$$\begin{aligned} \sigma_{MI} &= log_2 \frac{1000/n}{1000/n \cdot 1000/n} \\ &= log_2 (n/1\,000\,000) \end{aligned}$$

In Tabelle 3.1 sind die 50 stärksten Verbindungen aus 7000 Kollokationspaaren aufgeführt, die intellektuell als gute Kandidaten für statistische Kollokationen ausgewählt wurden. Um die Maße gut vergleichen zu können, sind die Werte der anderen beiden betrachteten Signifikanzmaße der Tabellen mit aufgeführt.

z-Score

Die Berechnung des Z-Scores ist eine spezielle Transformationsregel aus der Statistik. Die Häufigkeit des Kovorkommens zweier Wörter a und b, bezeichnet mit $H(a, b)$, bildet eine numerische Kenngröße, d. h. den Wert, den die Zufallsvariable X für das entsprechende Paar (a, b) annimmt. Zur Berechnung des z-Score-Maßes geht man davon aus, dass diese Zufallsvariable mit Erwartungswert μ und Varianz σ normalverteilt ist. μ und σ Der Erwartungswert wird durch den Mittelwert und die Varianz durch die Standardabweichung der untersuchten Stichprobe geschätzt. Wenn man diese Transformation auf alle Werte der Zufallsvariablen X einer Verteilung anwendet, erhält man eine neue Zufallsvariable, die um den Mittelwert 0 mit der Standardabweichung 1 normalverteilt ist. Diese $N(0, 1)$-verteilte Zufallsvariable ist das Signifikanzmaß σ_z:

$$\sigma_z = \frac{X - \mu}{\sigma} \tag{3.2}$$

Wort a	$H(a)$	Wort b	$H(b)$	$H(a,b)$	σ_{MI}	σ_{tani}	σ_{CBA}
Untiefe	14	belorussisch	1	1	19,0	0,071	6
Tycho	21	Brahe	22	14	17,8	0,483	70
Biermösl	40	Blosn	32	29	17,4	0,674	140
Ennio	45	Morricone	12	9	16,9	0,188	43
Wigald	45	Boning	62	42	16,8	0,646	195
Untiefe	14	Kuropaty	5	1	16,7	0,056	5
Programmiersprachen	26	Grundkonzepte	4	1	16,1	0,035	5
Hoffmanns	121	Lebensansichten	1	1	15,9	0,008	5
Caterina	56	Valente	25	11	15,8	0,157	49
Gleichung	151	Gleichung	151	159	15,7	1,112	683
Rialto	9	Wendlandt	20	1	15,3	0,036	5
Rothenburg	74	Tauber	85	34	15,3	0,272	143
Elster	51	Pleiße	13	3	15,1	0,049	13
Addis	214	Abeba	199	189	15,0	0,844	774
Meeresregion	2	Institutes	121	1	14,9	0,008	4
Spannen	42	Floaten	6	1	14,9	0,021	4
Vanity	22	Fair	196	17	14,9	0,085	70
Rosenkavalier	91	Marschallin	20	7	14,8	0,067	29
Kajo	101	Schommer	107	40	14,8	0,238	162
Rheinischer	14	Merkur	196	10	14,7	0,050	41
Addis	214	Abbeba	4	3	14,7	0,014	13
Regierungspräsident	130	Antwerpes	27	11	14,5	0,075	44
Bayernkurier	45	Scharnagl	44	5	14,2	0,059	20
Darius	205	Milhaud	31	16	14,2	0,073	62
Miriam	379	Makeba	14	13	14,2	0,034	51
Placido	75	Domingo	296	51	14,1	0,159	195
Jean-Michel	56	Jarre	8	1	14,0	0,016	4
Corriere	207	della	303	132	13,9	0,349	498
Lionel	197	Jospin	294	112	13,8	0,295	419
Samsung	153	Goldstar	14	4	13,8	0,025	16
Parthenon	16	Akropolis	68	2	13,8	0,024	8
Keith	321	Jarrett	46	27	13,7	0,079	101
Gaddafi	44	Muammar	53	4	13,7	0,043	15
Katers	14	Murr	43	1	13,6	0,018	4
Ezer	77	Weizmann	28	3	13,4	0,029	11
Alija	350	Izetbegovic	514	237	13,3	0,378	846
Miroslav	394	Vitous	4	2	13,2	0,005	8
Java	244	Programmiersprache	63	19	13,2	0,066	68
Juristische	86	Wochenschrift	39	4	13,1	0,033	15
Wim	334	Duisenberg	34	13	13,1	0,037	46
Ungeziefer	103	Samsa	9	1	13,0	0,009	4
Regensburger	700	Domspatzen	20	15	13,0	0,021	53
Java	244	Sumatra	65	14	12,7	0,048	48
Hammer	823	Sichel	111	80	12,7	0,094	272
Oedipus	35	Strawinsky	134	4	12,6	0,024	14
Luciano	331	Pavarotti	168	47	12,6	0,104	160
Blitz	379	Donner	181	57	12,6	0,113	193
Ignacio	173	Lopez	302	43	12,6	0,100	146
Tel	1268	Aviv	855	839	12,5	0,653	2796
Konditor	47	Praline	28	1	12,5	0,013	4

Tabelle 3.1.: Signifikante Kollokationspaare nach dem *Mutual Information Index*

Der z-Score eines Wertes gibt also die Größe und Richtung seiner Abweichung von der Verteilung der Zufallsgröße an. Im Vergleich zu anderen Signifikanzmaßen werden Verbindungen zu häufigen Wörtern überdurchschnittlich hoch bewertet, so dass aus der Liste potentieller Kandidaten Stopwörter wie z. B. *und* entfernt werden müssen.

Maximum-Likelihood-Verhältnis

Das Maximum-Likelihood-Prinzip geht in seinem Ansatz auf C. F. Gauss zurück und wurde in seiner heutigen Allgemeinheit von R. A. Fisher entwickelt. Eine Beschreibung findet sich beispielsweise in [Wt85, S. 31 ff]. Darauf basiert das *Maximum-Likelihood-Maß* σ_{ML}. Für die Ereignisse *Wort kommt vor* und *Wort kommt nicht vor*, ermittelt für jede Position im Text und für die beiden Wörter a und b, wird berechnet, ob das Verhältnis von Vorkommen und Nichtvorkommen von a unter der Bedingung, dass b vorkommt, signifikant anders ist als dieses Verhältnis unter der Bedingung, dass b nicht vorkommt (d. h. man untersucht, ob $P(a|b)/P(\neg a|b)$ einen signifikant anderen Wert als $P(a|\neg b)/P(\neg a|\neg b)$ annimmt). Das untersucht man durch Bildung einer Vierfeldertafel für die vier verschiedenen Ereignisse und Bildung deren Randsummen $P(a|b)+P(a|\neg b)$, $P(\neg a|b)+P(\neg a|\neg b)$, $P(a|b)+P(\neg a|b)$ und $P(a|\neg b) + P(\neg a|\neg b)$.

Tanimoto-Maß

Das Tanimoto-Maß, 1983 von Ozawa vorgestellt, fußt auf der Mengentheorie. Es gibt den Grad der Überlappung der Mengen der Beispielsätze an: Sei T_a die Menge der Sätze, die das Wort a enthalten, T_b die Menge der Sätze mit Wort b. Dann ist das Tanimoto-Maß bestimmt durch das Verhältnis der Anzahl der Sätze, die beide Wörter enthalten zur Anzahl der Sätze, in denen mindestens eines der Wörter auftaucht.

$$\sigma_{tani}(a,b) = \frac{|T_a \cap T_b|}{|T_a \bigcup T_b|} = \frac{H(a,b)}{H(a) + H(b) - H(a,b)} \tag{3.3}$$

σ_{tani} ergibt 1, wenn zwei Wörter immer zusammen im Satz stehen, und 0, wenn es im Korpus keinen Satz gibt, der beide Wörter enthält.

3.1.2. Das Common-Birthday-Maß

Im Wortschatz-Projekt wurde zur Berechnung der Kollokationen ein neues Signifikanzmaß hergeleitet, das sich an das von Yuval formulierte *Common-Birthday-*

Wort a	$H(a)$	Wort b	$H(b)$	$H(a,b)$	σ_{tani}	σ_{CBA}	σ_{MI}
Gleichung	151	Gleichung	151	159	1,112	683	15,7
Addis	214	Abeba	199	189	0,844	774	15,0
Biermösl	40	Blosn	32	29	0,674	140	17,4
Tel	1268	Aviv	855	839	0,653	2796	12,5
Wigald	45	Boning	62	42	0,646	195	16,8
Tycho	21	Brahe	22	14	0,483	70	17,8
Alija	350	Izetbegovic	514	237	0,378	846	13,3
Corriere	207	della	303	132	0,349	498	13,9
Berti	794	Vogts	1731	651	0,347	2031	11,8
Bill	2882	Clinton	4037	1772	0,344	4653	10,2
Kultusminister	1063	Zehetmair	1081	539	0,336	1680	11,8
Rio	1539	Janeiro	487	504	0,331	1650	12,3
Edmund	2213	Stoiber	3791	1419	0,309	3791	10,3
Lionel	197	Jospin	294	112	0,295	419	13,8
Rothenburg	74	Tauber	85	34	0,272	143	15,3
Slobodan	643	Milosevic	1522	462	0,271	1441	11,8
Oskar	1736	Lafontaine	2082	800	0,265	2231	10,7
Kajo	101	Schommer	107	40	0,238	162	14,8
Silvio	519	Berlusconi	1116	313	0,237	995	12,0
Felipe	425	Gonzalez	627	186	0,215	613	12,4
Heide	1283	Simonis	547	297	0,194	913	11,6
Angela	2051	Merkel	1104	501	0,189	1398	10,7
Ennio	45	Morricone	12	9	0,188	43	16,9
Bundeskanzler	3104	Kohl	8515	1791	0,182	4074	9,0
Willy	1429	Brandt	1031	377	0,181	1076	10,9
Placido	75	Domingo	296	51	0,159	195	14,1
Caterina	56	Valente	25	11	0,157	49	15,8
Hansa	350	Rostock	1436	241	0,156	754	11,8
Rupert	731	Murdoch	555	168	0,150	515	11,6
Björn	622	Engholm	381	131	0,150	419	12,0
Frankfurt	13044	Main	2358	1840	0,136	4086	8,8
Joschka	702	Fischer	4795	624	0,128	1693	10,4
Botho	233	Strauß	1576	195	0,121	619	12,0
tatenlos	232	zusehen	431	69	0,116	228	12,3
Blitz	379	Donner	181	57	0,113	193	12,6
Umweltministerin	525	Merkel	1104	155	0,105	446	11,0
Verkehrsminister	521	Wissmann	663	112	0,104	332	11,3
Luciano	331	Pavarotti	168	47	0,104	160	12,6
Ignacio	173	Lopez	302	43	0,100	146	12,6
Konstantin	920	Wecker	628	140	0,099	397	10,8
Pablo	282	Picasso	486	68	0,097	215	11,9
Romeo	343	Julia	1080	124	0,096	369	11,3
Steuergeldern	251	Verschwendung	373	54	0,095	174	12,1
Hammer	823	Sichel	111	80	0,094	272	12,7
Rolling	249	Stone	363	52	0,093	168	12,1
Vanity	22	Fair	196	17	0,085	70	14,9
Mario	2685	Basler	1291	307	0,084	735	9,4
Keith	321	Jarrett	46	27	0,079	101	13,7
Finanzminister	2957	Waigel	3966	486	0,075	1003	8,3
Regierungspräsident	130	Antwerpes	27	11	0,075	44	14,5

Tabelle 3.2.: Signifikante Kollokationspaare nach dem *Tanimoto-Maß*

Problem anlehnt.

Das Verfahren basiert auf der folgenden, aus der elementaren Wahrscheinlichkeitsrechnung bekannten Aufgabe[WW98, S. 97]: In einem Raum befinden sich 23 Schüler. Wie groß ist die Wahrscheinlichkeit, dass zwei dieser Schüler am gleichen Tag Geburtstag haben? (Zur allgemeinen Verwunderung stellt sich heraus, daß diese Wahrscheinlichkeit knapp über 50% liegt.) Wir verändern die Aufgabe zunächst leicht und übertragen sie dann auf das folgende Kollokationsproblem.

Gegeben sind zwei Wörter a und b. Wie groß ist die Wahrscheinlichkeit, dass unter n Sätzen $H(a,b)$ Stück sind, die beide Wörter a und b enthalten? Dazu sei bekannt, daß insgesamt $H(a)$ Sätze das Wort a und $H(b)$ Sätze das Wort b enthalten. Die gesuchte Wahrscheinlichkeit soll berechnet werden unter der zusätzlichen Annahme, daß die Auftreten von a und b Wörter unabhängig voneinander sind.

Dieses Problem stellt eine Variante des Common-Birthday-Problems dar: In einem Raum befinden sich $H(a)$ Jungen und $H(b)$ Mädchen. Wie groß ist die Wahrscheinlichkeit, dass es ein Paar (also ein Junge und ein Mädchen) gibt, das am gleichen Tag Geburtstag hat? Wie groß ist die Wahrscheinlichkeit, daß es $H(a,b)$ Paare gibt, die jeweils am gleichen Tag Geburtstag haben (d. h. wir erlauben für diese $H(a,b)$ Pärchen $H(a,b)$ verschiedene Geburtstage)?

Dabei soll zusätzlich angenommen werden, daß keine zwei Jungen und keine zwei Mädchen am gleichen Tag Geburtstag haben. Diese Annahme stellt keine wesentliche Einschränkung dar, wenn die Anzahl der Jungen und Mädchen sehr klein im Vergleich zur Anzahl der Tage eines Jahres ist.

Folgende Tabelle zeigt, wie wir das Kollokationsproblem in die beschriebene Variante des Common-Birthday-Problems überführen können:

Symbol	Common-Birthday-Problem	Kollokationsproblem
$H(a)$	Anzahl der Jungen	Anzahl der Sätze, die das Wort a enthalten
$H(b)$	Anzahl der Mädchen	Anzahl der Sätze, die das Wort b enthalten
n	Anzahl der Tage im Jahr	Gesamtzahl aller Sätze
$H(a,b)$	Anzahl der Paare mit gemeinsamen Geburtstag	Anzahl der Sätze, die beide Wörter a und b enthalten

Da die Gesamtzahl n aller Sätze stets groß gegen die Anzahlen $H(a)$ und $H(b)$ sein wird (typischerweise mindestens um den Faktor 1000), ist die zusätzliche Annahme über die Verschiedenheit der Geburtstage innerhalb der Jungen bzw. Mädchen gerechtfertigt. Zum besseren Verständnis werden die folgenden Rechnungen immer im Kontext des Common-Birthday-Problems beschrieben.

Die Wahrscheinlichkeit, daß von $H(a)$ Jungen und $H(b)$ Mädchen kein Paar am gleichen Tag Geburtstag hat, ist:

$$p_0 = \frac{n - H(a)}{n} \cdot \frac{n - H(a) - 1}{n - 1} \cdot \ldots \cdot \frac{n - (H(a) + H(b)) + 1}{n - H(b) + 1}$$

Um das Ereignis des gemeinsamen Geburtstages eines Mädchens mit einem Jungen zu vermeiden, bleiben für das erste Mädchen $n - H(a)$ mögliche Geburtstage von insgesamt n, für das zweite Mädchen noch $n - H(a) - 1$ mögliche Geburtstage usw.

Betrachten wir nun den Fall von genau einem Paar mit gemeinsamen Geburtstag. Die Wahrscheinlichkeit hierfür beträgt

$$p_1 = H(b) \cdot \frac{H(a)}{n} \cdot \frac{n - H(a)}{n - 1} \cdot \frac{n - H(a) - 1}{n - 2} \cdot \ldots \cdot \frac{n - (H(a) + H(b)) + 2}{n - H(b) + 1}$$

Die einzelnen Faktoren resultieren daher, dass wir jedes von den $H(b)$ Mädchen für das Paar auswählen können, so dass dieses Mädchen bei $H(a)$ von n möglichen Geburtstagen auf einen der Jungen trifft, aber die anderen Mädchen analog oben die bisher vergebenen Geburtstage vermeiden müssen, um keine weiteren Paare zu bilden.

Betrachten wir nun den Fall von genau zwei Paaren mit jeweils gemeinsamen Geburtstag. Die Wahrscheinlichkeit hierfür beträgt

$$p_2 = \binom{H(b)}{2} \cdot \frac{H(a)}{n} \cdot \frac{H(a) - 1}{n - 1} \cdot \frac{n - H(a)}{n - 2} \cdot \frac{n - H(a) - 1}{n - 3} \cdot \ldots \cdot \frac{n - (H(a) + H(b)) + 3}{n - H(b) + 1}$$

Die Faktoren resultieren wieder daher, daß wir auf $\binom{H(b)}{2}$ Arten zwei von den $H(b)$ Mädchen für die Paare auswählen können, so dass das erste Mädchen bei $H(a)$ von n möglichen Geburtstagen auf einen der Jungen trifft, das zweite bei verbleibenden $H(a) - 1$ von $n - 1$ Tagen, die anderen Mädchen aber analog oben die bisher vergebenen Geburtstage vermeiden müssen, um keine weiteren Paare zu bilden.

Allgemein erhält man für $H(a, b)$ Paare:

$$\begin{aligned} p_{H(a,b)} \quad = \quad & \binom{H(b)}{H(a,b)} \cdot \frac{H(a)}{n} \cdot \frac{H(a) - 1}{n - 1} \cdot \ldots \cdot \frac{H(a) - H(a,b) + 1}{n - H(a,b) + 1} \cdot \frac{n - H(a) - 1}{n - H(a,b)} \cdot \\ & \ldots \cdot \frac{n - (H(a) + H(b)) + H(a,b) + 1}{n - H(b) + 1} \end{aligned}$$

Da $H(a)$ und $H(b)$ im Vergleich zu n klein sind, unterscheiden sich in den einzelnen Gruppen von Faktoren die benachbarten Glieder nur wenig, so dass die folgende

Approximation für uns ausreichend gut ist:

$$p_{H(a,b)} \approx \frac{1}{H(a,b)!} \cdot H(b)^{H(a,b)} \cdot \left(\frac{H(a)}{n}\right)^{H(a,b)} \cdot \left(\frac{n - H(a)}{n}\right)^{H(b)} \tag{3.4}$$

Setzen wir weiter $x = ab/n$, so gilt

$$\left(\frac{n - H(a)}{n}\right)^{H(b)} \approx e^{-x},$$

also schließlich

$$p_{H(a,b)} \approx \frac{1}{H(a,b)!} \cdot x^{H(a,b)} \cdot e^{-x}. \tag{3.5}$$

Im folgenden soll mit dieser Approximation weitergerechnet werden. Uns interessiert die Wahrscheinlichkeit, daß mindestens $H(a,b)$ Paare auftreten. Die Wahrscheinlichkeit $q_{H(a,b)}$ dafür beträgt offensichtlich

$$q_{H(a,b)} = \sum_{i=H(a,b)}^{\infty} p_i = e^{-x} \sum_{i=H(a,b)}^{\infty} \frac{1}{i!} \cdot x^{H(a,b)}.$$

Nehmen die Summanden in der obigen Summe schnell genug ab, so reicht es, nur den ersten Summanden zu betrachten. Wenn wir insgesamt Abweichungen von 10% akzeptieren wollen, ist dafür die Bedingung $(H(a,b)+1)/x < 0,1$ ausreichend. In den anderen Fällen benutzen wir wegen $\sum p_i = 1$ die Formel

$$q_{H(a,b)} = \sum_{i=0}^{H(a,b)-1} p_i = e^{-x} \sum_{i=0}^{H(a,b)-1} \frac{1}{i!} \cdot x^{H(a,b)}. \tag{3.6}$$

Definition 3.1 (Common-Birthday-Maß) *Als **Signifikanz** $\sigma_{CBA}(A,B)$ für das gemeinsame Auftreten der Wörter a und b definieren wir den negativen Logarithmus der obigen Wahrscheinlichkeit. Damit ergeben sich für die Signifikanz die folgenden Formeln:*

Sei $H(a)$ die Anzahl der Sätze mit Wort a, $H(b)$ die Anzahl der Sätze mit Wort b, n die Gesamtanzahl aller Sätze und $H(a,b)$ die Anzahl der Sätze, welche die Wörter a und b enthalten. Wir setzen $x = H(a)H(b)/n$, und definieren:

1. *Gilt $(H(a,b)+1)/x < 0.1$ (dies ist der typische Fall), so setzen wir*

$$\sigma_{CBA}(a,b) = x - \log_{10}\left(\sum_{i=0}^{H(a,b)-1} \frac{1}{i!} \cdot x^{H(a,b)}\right) \tag{3.7}$$

2. Anderenfalls setzen wir

$$\sigma_{CBA}(a,b) = 1/2\,(x \cdot \log_{10} e - H(a,b) \cdot \log x + log_{10}\,(H(a,b)!)) \tag{3.8}$$

Zwei Wörter a und b sind dann signifikante Kollokationen, wenn das Signifikanzmaß $\sigma_{CBA}(a,b) \geq 4$ ist. Diese Schwelle wurde nach Inaugenscheinnahme einiger berechneter Kollokationen festgelegt.

Bemerkenswert an der Definition ist, daß die Signifikanz nicht nur von den relativen Häufigkeiten $H(a)/n, H(b)/n$ und $H(a,b)/n$ abhängt, sondern die Signifikanz bei konstanten relativen Häufigkeiten zusammen mit der Korpusgröße wächst. Anschaulich läßt sich dieser Effekt damit erklären, dass uns das einmalige gemeinsame Auftreten zweier Wörter (z. B. *Katze* und *Sack*) keine Information gibt, doch das wiederholte Auftreten beider Wörter in einem entsprechend größeren Korpus uns einen Zusammenhang zwischen den Wörtern vermuten lässt (hier gegeben durch die Redewendungen *„die Katze im Sack kaufen“* und *„die Katze aus dem Sack lassen“*).

Common-Birthday-Maß auf Satzebene

Mit dieser Methode werden sowohl syntaktisch-semantische Kollokationen als auch andere signifikant häufige Wortpaare erkannt, soweit das im Rahmen des Textkorpus möglich ist. Dazu gehören zahlreiche Head-Modifier-Relationen (Relationspaare aus der Dependenzgrammatik nach Tesnière, z. B. (*beißt – Hund*)).

Speziell lassen sich hier auch Namen geographischer Orte finden, die räumlich benachbart und politisch ähnlich bedeutsam sind. Weiterhin können durch nichtdeutsche Stoppworte viele Wörter ebendieser Sprache gefunden werden (englisch, spanisch, bayerisch).

Common-Birthday-Maß auf Nachbarebene

Das oben beschriebene Signifikanzmaß wurde auch auf der Basis von Wortnachbarn berechnet, je einmal für die linken und rechten Nachbarn des jeweiligen Wortes. Die Berechnung entspricht der von σ_{CBA} nach Definition 3.1. Zur Berechnung des Signifikanzmaßes auf Basis der linken Nachbarn, σ_{CBA_nbli}, verwenden wir in den Formeln 3.7 und 3.8 als Wert für $H(a,b)$ die Anzahl der Sätze, in denen das Wort a auf das Wort b folgt, in denen also b links von a steht. Analog verwenden wir zur Berechnung der rechten Kollokationen nach dem Maß σ_{CBA_nbre} für $H(a,b)$ die Anzahl der Sätze, in denen das Wort a direkt vor dem Wort b steht und erhalten somit alle Wörter b, die signifikant häufig rechts von a auftauchen.

Wort a	$H(a)$	Wort b	$H(b)$	$H(a,b)$	σ_{CBA}	σ_{tani}	σ_{MI}
Bill	2882	Clinton	4037	1772	4653	0,344	10,2
Frankfurt	13044	Main	2358	1840	4086	0,136	8,8
Bundeskanzler	3104	Kohl	8515	1791	4074	0,182	9,0
Edmund	2213	Stoiber	3791	1419	3791	0,309	10,3
Tel	1268	Aviv	855	839	2796	0,653	12,5
Oskar	1736	Lafontaine	2082	800	2231	0,265	10,7
Berti	794	Vogts	1731	651	2031	0,347	11,8
Joschka	702	Fischer	4795	624	1693	0,128	10,4
Kultusminister	1063	Zehetmair	1081	539	1680	0,336	11,8
Rio	1539	Janeiro	487	504	1650	0,331	12,3
Gerhard	9696	Schröder	2943	852	1637	0,072	7,8
Präsident	18882	Clinton	5019	1058	1583	0,046	6,4
wies	4152	darauf	22839	996	1465	0,038	6,3
Slobodan	643	Milosevic	1522	462	1441	0,271	11,8
Angela	2051	Merkel	1104	501	1398	0,189	10,7
Franz	10540	Beckenbauer	1332	603	1254	0,053	8,3
Willy	1429	Brandt	1031	377	1076	0,181	10,9
Frankfurt	13044	Eintracht	1178	514	1013	0,037	8,0
Finanzminister	2957	Waigel	3966	486	1003	0,075	8,3
Silvio	519	Berlusconi	1116	313	995	0,237	12,0
Heide	1283	Simonis	547	297	913	0,194	11,6
Alija	350	Izetbegovic	514	237	846	0,378	13,3
in	1283649	Hannover	5522	3368	811	0,003	1,8
Addis	214	Abeba	199	189	774	0,844	15,0
Hansa	350	Rostock	1436	241	754	0,156	11,8
Kurt	5411	Biedenkopf	745	321	754	0,055	9,2
Mario	2685	Basler	1291	307	735	0,084	9,4
Gleichung	151	Gleichung	151	159	683	1,112	15,7
Botho	233	Strauß	1576	195	619	0,121	12,0
Felipe	425	Gonzalez	627	186	613	0,215	12,4
in	1283649	Leipzig	3339	2093	523	0,002	1,9
Rupert	731	Murdoch	555	168	515	0,150	11,6
Corriere	207	della	303	132	498	0,349	13,9
Umweltministerin	525	Merkel	1104	155	446	0,105	11,0
Kaffee	2113	Kuchen	597	164	421	0,064	9,9
Lionel	197	Jospin	294	112	419	0,295	13,8
Björn	622	Engholm	381	131	419	0,150	12,0
Konstantin	920	Wecker	628	140	397	0,099	10,8
Bundeskanzler	3104	Vranitzky	456	158	395	0,046	9,7
Landeshauptstadt	1363	München	22094	269	378	0,012	6,1
Romeo	343	Julia	1080	124	369	0,096	11,3
Sache	10451	eigener	3384	266	354	0,020	5,8
Reinhard	2930	Höppner	443	142	354	0,044	9,7
Lafontaine	2082	SPD	10361	226	333	0,018	6,3
Verkehrsminister	521	Wissmann	663	112	332	0,104	11,3
Joseph	1983	Beuys	342	122	330	0,055	10,4
Hans	15376	Eichel	357	169	328	0,011	7,9
Umweltministerin	525	Angela	2051	122	306	0,050	9,7
Kurt	5411	Beck	1436	160	283	0,024	7,3
essen	1571	trinken	1184	121	274	0,046	8,9

Tabelle 3.4.: Signifikante Kollokationspaare nach dem *Common-Birthday-Maß*

Im wesentlichen sind die gefundenen Paare auch Kollokationen auf Satzebene, jedoch sind sie hier ganz anders gewichtet, und diejenigen Relationen treten stärker hervor, die sich in Strukturen benachbarter Wörter finden:

- Aufzählungen (wie Bundesländer), falls oft mehr als zwei Objekte genannt werden
- Mehrwortbegriffe, Personennamen, Titel von Personen (akademische Titel oder Berufe wie Gesundheitsminister, Regisseur, ...)
- Eigenschaften (Adjektive, die zur näheren Beschreibung vor dem Wort stehen; das sind aber nicht unbedingt typische, beschreibende Eigenschaften (*der schwere Amboss*), sondern Eigenschaften, die zur näheren Klassifikation eines bestimmten Objektes dienen)
- Head-Modifier-Strukturen (*Hund – bellt*)

3.1.3. Schnitt zweier Kollokationsmengen

Unter den Kollokationen eines Wortes finden sich Relationen verschiedenartiger Natur. Wörter der gleichen Klasse (wie Wochentage oder Farben) zählen ebenso dazu wie Wörter aus Head-Modifier-Strukturen oder Synonyme.

Beispiel *Sonntag*:
Wörter der gleichen Klasse: *Samstag, Freitag, Montag, ...*
Wörter aus Head-Modifier-Strukturen: *kommenden, vergangenen, verkaufsoffenen, autofreier, ...*
weitere Kollokationen: *Uhr, Nacht, Stichwahl, Gasteig, Tatort, ausgeschlafen*

Außerdem sind bei Homonymen die Kollokationen der verschiedenen Wortbedeutungen gemischt. Ein anderes Wort aus einer gleichen Klasse weist als Kollokationen auch viele andere Vertreter dieser Klasse auf, aber keine Wörter aus Klassen, in denen nur das ursprüngliche Wort enthalten ist.

Beispiel:
Reis ist sowohl eine Hülsenfrucht als auch der Name eines Fußballers, unter den Kollokationen finden sich also unter anderem Hülsenfrüchte und Fußballer (Tabelle 3.5).

Greift man sich nun aus diesen Kollokationen den Namen eines Fußballers oder der Fußballmannschaft heraus und bildet die Schnittmenge der Kollokationen von *Reis* und beispielsweise *Bochum*, erhält man in Tabelle 3.6 die Namen der anderen Mitspieler des VfL Bochum (und außerdem die Stadt *Essen*). Als Maß für die Gewichtung der Kollokationen bietet sich die Summe der Signifikanzmaße $\sigma_{CBA}(\text{Reis}, i) + \sigma_{CBA}(\text{Bochum}, i)$ an.

Wort	Wert	Wort	Wert	Wort	Wert
Bohnen	54	Weizen	26	Japan	10
Gospodarek	50	Zucker	24	Maniok	10
Baluszynski	47	Gudjonsson	23	Gramm	10
Tonnen	43	angebaut	23	Fleisch	10
Waldoch	40	Kartoffeln	22	Angra	10
Kracht	38	Michalke	22	essen	9
Wosz	38	Jack	21	Pjöngjang	9
Közle	37	Hutwelker	18	Öl	9
Stickroth	37	Hirse	17	Speiseöl	9
Mamic	34	Südkorea	17	Tee	9
Mais	33	Fisch	16	Okocha	8
Nordkorea	30	Bananen	15	Handvoll	8
Zuckerrohr	28	Mehl	15	Bindewald	8
Tapalovic	28	Nudeln	14	Cabrita	8
Peschel	27	Gemüse	13	Grundnahrungsmittel	8
Donkow	27	Lieferung	12	Kaffee	8
Bochum	27	Pfund	11	⋮	⋮

Tabelle 3.5.: Kollokationen für *Reis*

Die Aspekte von *Reis* als Nahrungsmittel kann man weiter aufteilen in die Bereiche Reis als Feldfrucht, Reis als Produkt in Agrarstaaten oder einfach Reis als Grundnahrungsmittel. Für den ersten Bereich wurde *Weizen* als weiterer Vertreter der Klasse *Feldfrüchte* gewählt. Da den Relationen keine Syntaxanalyse zu Grunde liegt, finden sich unter den Kollokationen auch Wörter, die signifikant häufig im Bereich der Feldfrüchte auftauchen wie *Tonnen* oder *angebaut*. Analog ist die zweite Tabelle aufgebaut, für die *Tee* als weiterer Klassenvertreter gewählt wurde. Die niedrige Signifikanz der Kollokationen mit *Obst* überträgt sich aus den Einzelrelationen auf die Schnittmenge.

In der dritten Tabelle wurde der Oberbegriff der Klasse, *Grundnahrungsmittel,* als

Reis – Bochum

Wort	Wert	Wort	Wert	Wort	Wert
Wosz	120	Peschel	75	Hutwelker	44
Baluszynski	115	Michalke	72	Jack	44
Gospodarek	113	Mamic	71	Eberl	31
Waldoch	112	Donkow	61	Essen	20
Közle	86	Gudjonsson	57	Schreiber	15
Kracht	80	Tapalovic	55	Winkler	11
Stickroth	78				

Tabelle 3.6.: Schnittmenge der Kollokationen von *Reis* mit denen von *Bochum*

Reis – Weizen

Wort	Wert
Mais	69
Tonnen	65
Kartoffeln	45
angebaut	35
Zuckerrohr	33
Hirse	27
Getreide	12
Baumwolle	12
Sojabohnen	10
Obst	8

Reis – Tee

Wort	Wert
Kaffee	80
Zucker	36
angebaut	27
Brot	21
Bananen	19
Wasser	18
Baumwolle	16
Tabak	14
Kilo	9
Obst	8

Reis – Grundnahrungsmittel

Wort	Wert
Zucker	28
Kartoffeln	26
Mehl	20
Brot	8

Tabelle 3.7.: Schnittmenge der Kollokationen von *Reis* mit denen von *Weizen, Tee* und *Grundnahrungsmittel*

zweiter Begriff für die Schnittmengenbildung gewählt. In diesem Fall wurden die anderen Vertreter der Klasse gefunden, da in den verwendeten Textkorpora Sätze der Art

> Grundnahrungsmittel wie Mehl, Reis oder Zucker zum Gelieren brachten sie in den Südosten; Fertigsuppen, Margarine, Nudeln und andere sättigende Lebensmittel, die schnell zubereitet werden können. (Quelle: *Frankf. Rundschau 1992*)
> Die bereits fünfwöchige Trockenheit hat die Mais- und Bohnenaussaat – beides Grundnahrungsmittel der Bevölkerung – empfindlich geschädigt. (Quelle: *TAZ 1987*)

auftauchen. In vielen Fällen kann aber nicht davon ausgegangen werden, dass sich der Klassenoberbegriff unter den Kollokationen befindet, wohingegen die Vertreter einer Klasse meist als Cluster gefunden werden.

Neben Kohyponymen kann man durch Bildung der Schnittmenge der Kollokationen auch Wörter nach anderen lexikalischen Funktionen finden. So tauchen z. B. unter den Kollokationen von *Oberbürgermeister* die Namen von Oberbürgermeistern bekannter Städte auf. Unter den Kollokationen der Städte finden sich auch Namen von für diese Stadt bedeutsamen Persönlichkeiten, sei es historisch, kulturell oder tagespolitisch. In der Schnittmenge wird man die Oberbürgermeister der Städte finden, wenn auch nur so aktuell, wie es die eingelesenen Texte ermöglichen. Auf die gleiche Art finden sich Hauptstädte, Romanautoren, Objekte von Handlungen usw.

3.2. Extraktion von Konzepten aus Kollokationen durch Verwendung von Wortvektoren

3.2.1. Gemeinsame Kollokationen und Nachbarn

Bisher wurde die automatische Ermittlung von signifikanten Kollokationen untersucht, indem verschiedene Signifikanzmaße auf Wörter in einer bestimmten Umgebung angewandt wurden. Darauf aufbauend kann man semantisch verwandte, also im gleichen Kontext verwendete Wörter bestimmen, indem man die Kollokationen der Wörter und deren Signifikanz miteinander vergleicht.

Einen möglichen Ansatz zu diesem Vergleich bietet die Vektoranalyse [SG83]. Die Menge der Kollokationen eines Wortes kann als Vektor im n-dimensionalen Raum betrachtet werden. Diesen Vektor bezeichnen wir als *Kollokationsvektor* $\vec{k}$. Die i-te Spalte des Kollokationsvektors eines Wortes a ist mit dem Wert x besetzt, wenn der Signifikanzwert eines Kollokationsmaßes zwischen dem Wort a und dem i-ten Wort des Gesamtwortschatzes gleich x ist.

Beispiel:
Bestehe der gesamte betrachtete Wortschatz aus folgenden Wörtern:

Nr.	Wort
1	Bär
2	beißt
3	dicker
4	Hund
5	Mann

Zwischen diesen Wörtern bestehen folgende (symmetrischen) Kollokationsmaße:
$\sigma_{CBA}(\text{Bär, Hund}) = 5, \quad \sigma_{CBA}(\text{beißt, Hund}) = 22,$
$\sigma_{CBA}(\text{beißt, Mann}) = 8, \quad \sigma_{CBA}(\text{dicker, Hund}) = 4,$
$\sigma_{CBA}(\text{dicker, Mann}) = 6, \quad \sigma_{CBA}(\text{Hund, Mann}) = 22$
Dann sind die Kollokationsvektoren von *Bär, Hund* und *Mann*:
$\vec{k}(\text{Bär}) = (0, 0, 0, 5, 0),$
$\vec{k}(\text{Hund}) = (5, 22, 4, 0, 22),$
$\vec{k}(\text{Mann}) = (0, 8, 6, 22, 0)$

Zwei Wörter a, b kann man nun vergleichen, indem man das Skalarprodukt ihrer Kollokationsvektoren berechnet. Zur besseren Vergleichbarkeit mit dem Ausgangsmaß wird aus dem Skalarprodukt die Wurzel gezogen.

$$\sigma_{2prod}(a,b) = \sqrt{\sum_i (\sigma_{CBA}(a,i) \cdot \sigma_{CBA}(b,i))} \tag{3.9}$$

Im obigen Beispiel erhält man dann
σ_{2prod}(Hund, Mann) $= \sqrt{5 \cdot 0 + 22 \cdot 8 + 4 \cdot 6 + 0 \cdot 22 + 22 \cdot 0} \approx 14$,
σ_{2prod}(Bär, Hund) $= 0$,
σ_{2prod}(Bär, Mann) ≈ 10

Das Hauptaugenmerk bei diesem Signifikanzmaß liegt auf der Verwendung zweier Wörter **mit dem** gleichen Kontext. Die direkte Beziehung, d. h. die Verwendung zweier Wörter **im** gleichen Kontext, soll dem gegenüber zurückgestellt werden. Deshalb wird bei der Berechnung davon ausgegangen, dass ein Wort zu sich selbst keine Kollokation ist, also $\sigma_{CBA}(a,a) = 0$ ist. Nicht signifikante Kollokationen, das sind solche mit einen Signifikanzwert kleiner als 4, gehen ebenfalls mit dem Wert Null in die Berechnung ein, da die Speicherung dieser Relationen in unverhältnismäßig hohem Aufwand zum geringen Fehler bei der Berechnung stünde.

Bei der Berechnung der Kollokationen nach dem Common-Birthday-Maß konnten wir uns auf Wörter beschränken, die im gleichen Satz wie das Ausgangswort auftraten[1]. Das Signifikanzmaß σ_{2prod} ist nicht mehr auf Wörter in einem Wortfenster beschränkt, sondern berechnet die *Kollokationen zweiter Ordnung* zu einem Wort. Damit bezeichnen wir alle statistisch berechneten Relationen zwischen Wörtern, die auf Grund der Kollokationen der Wörter (statt nur auf Grund der Wortfrequenzen) berechnet werden.

Mögliche Kandidaten für diese Relation sind alle Wörter, die mit dem Ausgangswort gemeinsame Kollokationen besitzen. Deshalb werden bei der Berechnung der Kollokationen zweiter Ordnung zu einem Wort a zunächst die Kollokationen i mit den zugehörigen Signifikanzwerten $\sigma_{CBA}(a,i)$ gespeichert. Aus allen Kollokationen der Wörter i wird nun die Vereinigungsmenge aller Wörter b gebildet. Anschließend wird für alle Wörter b das Maß $\sigma_{2prod}(a,b)$ berechnet. Die Paare, für die das Maß eine gewisse Schwelle überschreitet, werden in der Datenbank gespeichert. Als sinnvolle Grenze wählten wir $\sigma_{2prod}(a,b) > 12$.

Die Berechnung der Kollokationen erfolgt mit dem Programm `sig_vec`, das eine Liste von `INSERT`-Statements zur Aufnahme in die Datenbank generiert. Dem Programm werden als Argumente die erste und letzte Wortnummer des Bereiches übergeben, für die die Kollokationen zweiter Ordnung berechnet werden sollen. Alternativ ist geplant, die Kollokationen für einzelne Wörter nach Bedarf von dem Programm berechnen zu lassen, dass für die WWW-Schnittstelle die Wortschatz-Daten anzeigt.

[1] Stoppwörter wurden von der Berechnung ausgeschlossen.

einstufig		zweistufig (Minimum)		zweistufig (Produkt)	
Wert	Wort	Wert	Wort	Wert	Wort
61	Katze	61	Sack	64	beißen
30	Herrchen	30	Frauchen	58	Tier
24	Herr	30	Maus	54	Hunde
22	Mann	29	Schwanz	53	Frauchen
22	beißt	29	Hunde	52	bellten
22	Tier	24	Knecht	50	beiße
19	Frau	24	Herrn	49	Maus
18	Schwanz	24	beißen	44	Schwanz
17	Leine	23	Doktor	42	belle
17	Hundehalter	22	beiße	39	Hans-Jochen
16	Katz	22	hineinbeißt	38	Katzen
15	Mensch	22	hineinbeißen	35	bellen
14	bellt	22	ausbeißt	34	tot
14	bellte	22	festbeißen	33	Herrn
14	gekommen	22	abbeißt	33	spazierenführen
14	Kind	22	ausbeißen	32	spazierenfahren
12	bellen	22	abbeißen	31	hineinbeißt
11	geprügelter	22	entzwei	31	hineinbeißen
11	Katzen	22	Katze	31	ausbeißt
11	entlaufener	21	tot	31	beißt
11	gebissen	21	Dame	31	spaziere
10	Auto	21	Vierbeiner	30	festbeißen
10	harter	21	bellen	30	abbeißt
9	scho	20	zubeißt	30	spaziert
9	bunter	20	anbeißt	30	spazieren
8	dog	19	beißt	30	rennen
8	Esel	19	Tier	29	ausbeißen
8	Pfoten	19	bellten	29	abbeißen
7	Tierarzt	18	lieber	29	Heynckes
7	Zwillinge	18	kurzen	28	entzwei
7	begraben	18	belle	28	bellt
7	Kaninchen	18	Lieber	28	Herrchen
7	blasse	18	zusammenbeißen	27	Katze
7	armer	18	anbeißen	27	Vierbeiner
7	Reiseapotheke	17	zubeißen	27	Schmadtke
7	Gromit	16	Hause	27	Fell
7	Frauchen	16	bitte	27	Schrödingers
6	Sirius	16	Wesen	26	Zeyer
6	Rasse	15	Meerschweinchen	26	Spanring
6	Hause	15	Fell	26	kleines
6	spazieren	15	Geburt	25	Heidenreich
6	gebellt	15	junger	25	junger
6	Spaziergängers	15	Ihrem	25	Mäuse
6	Gassi	15	eigenes	25	taube

Tabelle 3.8.: Kollokationen für *Hund*

Als Basis bei der Berechnung kann man neben den Kollokationen auf Satzbasis auch die Kollokationen zwischen Wortnachbarn verwenden. Die Wörter mit gemeinsamen rechten bzw. linken Nachbarn erhält man mit den folgenden Funktionen:

$$\sigma_{2rechts}(a,b) = \sqrt{\sum_i (\sigma_{CBA\_nbre}(a,i) \cdot \sigma_{CBA\_nbre}(b,i))} \tag{3.10}$$

$$\sigma_{2links}(a,b) = \sqrt{\sum_i (\sigma_{CBA\_nbli}(i,a) \cdot \sigma_{CBA\_nbli}(i,a))} \tag{3.11}$$

Durch die Verwendung des Skalarproduktes bei der Berechnung der Kolllokationen zweiter Ordnung werden Relationen zwischen Wörtern a und b auch dann hoch bewertet, wenn nur eine der Verbindungen von a *oder* b zu den gemeinsamen Kollokationen stark ist. Um diesen Effekt zu vermindern, untersuchten wir auch folgendes Maß:

$$\sigma_{2min}(a,b) = \sqrt{\sum_i min(\sigma_{CBA}(a,i), \sigma_{CBA}(b,i))^2} \tag{3.12}$$

Hier wird aus jeder „Dimension“ nur die minimale Komponente zur Berechnung herangezogen und über diesen analog das Skalarprodukt berechnet.

Beispiel:
Die Wörter *Sonntag* und *Polizeigewahrsam* haben nur eine gemeinsame Kollokation: *Freitag* mit den Signifikanzwerten
σ_{CBA}(Sonntag, Freitag) = 100 und
σ_{CBA}(Polizeigewahrsam, Freitag) = 4.
σ_{2prod} ergibt für das Paar (*Sonntag – Polizeigewahrsam*) einen recht hohen Wert von $\sqrt{100 \cdot 4} = 20$; σ_{2min} hingegen ergibt den Wert $\sqrt{min(4,4)^2} = 4$. Daran sieht man, dass das Maß σ_{2prod} Relationen stark überbewertet, wenn nur eine der Werte nach σ_{CBA} sehr hoch, der andere jedoch kaum signifikant ist. Nichtsdestotrotz sollte dabei nicht vergessen werden, dass unter den Kollokationen zweiter Ordnung von *Sonntag Polizeigewahrsam* in der Rangfolge weit hinten steht.

Die Werte variieren, wenn sich beide Teil-Relationen der zweistufigen Kollokationen weniger stark unterscheiden. Angenommen, die Wörter *Nacht* und *Sonntag* hätten wieder nur die gemeinsame Kollokation *Freitag*. Hier sind die Signifikanzwerte:
σ_{CBA}(Sonntag, Freitag) = 100 und
σ_{CBA}(Nacht, Freitag) = 173.
Damit erhalten wir für σ_{2prod} rund 132 und $\sigma_{2min} = 100$. Die Kollokation ist also nach beiden Maßen signifikant.

Resultate:
Durch diese Signifikanzmaße und deren Kombinationen erhält man verstärkt semantische Cluster und Synonyme. So findet man z. B.:

- Kohyponyme (Wörter mit gleichem Oberbegriff) durch Anwendung des Maßes $\sigma_{2rechts}$
- Synonyme durch Kombination der Maße $\sigma_{2rechts}$ und $\sigma_{2rechts}$ zu $\sigma_{2rechts} \cdot \sigma_{2rechts}$

Ausgangswort	beste Kollokationen und Wert nach dem Maß $\sigma_{2links} \cdot \sigma_{2rechts}$
• Arzt	Professor (6567)
• Computer	Rechner (1260), Auto (1156), IBM (957), Computern (864)
• DM	Dollar (460902), Yen (46920), US-Dollar (33488), Pfund (28561), Franc (23100), Tonnen (23072), Lire (22270), Franken (21922), Stunden (15930), Gulden (11400), Pfennig (9918), Rubel (8883), Jahre (8633), Liter (8384), Kilowattstunden (8062), Schilling (7872), Francs (6885), ECU (6278), Ecu (5808), ...
• Duma	Parlament (330)
• erklärt	weiß (14160), heißt (13981), erklären (12880), erklärte (9720), gilt (9666), gesagt (5100), meint (4921), wußte (3390), erzählt (1989), dafür (1927), genau (1696), fragte (1674), betont (1562), ...
• FC	Fortuna (9570), HC (7125), SC (6216), Staatsanwaltschaft (4845), Babbel (4784), Helmer (4238), AC (3955), ...
• Insel	Inseln (1400), Ausstellung (1050), Tageszeitung (682), Zeitung (666), Provinz (585), Hauptinsel (392)
• Klinik	Krankenhaus (7350), Krankenhäuser (1474), Kliniken (1269), Flughafen (1003), Anstalt (975), Krankenhauses (672), Dienst (648)
• Krankheit	Verletzungen (1648), Unfall (910), Kosten (480), Virus (306), Mißhandlungen (299), Brandverletzungen (276), Schußverletzung (156)
• liest	schreibt (1794)
• Minister	Innenminister (21762), Herr (18384), Ministerpräsident (18032), Wirtschaftsminister (16650), Finanzminister (16571), Trainer (14353), Außenminister (10098), Verteidigungsminister (9912), Gesundheitsminister (7370), FDP-Vorsitzende (7332), Kultusminister (6528), CSU-Vorsitzende (6141), Professor (5980), Bürgermeister (5424), Regierungschef (5355), Umweltminister (5292), SPD-Vorsitzende (5250), Staatsanwalt (5222), Landwirtschaftsminister (4860), Arbeitsminister (4352), Staatsminister (4176), ...
• Müller	Kinkel (2610), Schmidt (1748), Bremer (1656), Bauer (1380), ...
• Regierungschef	Ministerpräsident (102 424), Außenminister (20313), Ministerpräsidenten (16878), Premier (14416), Premierminister (13965), Bürgermeister (11388), Innenminister (10224), Papst (9843), Finanzminister (8979), Chef 7812), ...
• Richter	Trainer (9504), Bürgermeister (6912), Stadtrat (6510), Professor (6237), Geschäftsführer (5073), Anwalt (4170), Oberbürgermeister (4152), Staatsanwalt (3570), Händler (3293), ...

Ausgangswort	beste Kollokationen und Wert nach dem Maß $\sigma_{2links} \cdot \sigma_{2rechts}$
• Rücktritt	Fall (4920), Verkauf (4699), Widerstand (1216), Schritt (1207), Waffenstillstand (1008), Rückzug (960), ...
• sagte	erklärte (70060), heißt (49446), betonte (42020), gilt (39278), forderte (36192), meinte (32318), kündigte (26571), warf (26460), weiß (25935), meint (25530), teilte (22876), erklärt (22792), hält (20979), ...
• spielt	spielen (71400), besteht (16272), sieht (11946), stehen (7293), zieht (4991), ...
• später	pro (52500), lang (17810), danach (5016), jetzt (4921), Ende (3393), kürzlich (3164), Anfang (2782), nun (2405), dazu (2121), ...
• Tor	Ergebnis (12035), Publikum (5232), Fenster (1710), Treffer (1344), Weg (936), ...
• Torwart	Trainer (5978), Oberbürgermeister (5141), Manager (4032), Kapitän (2312), Stürmer (1176), ...
• trinken	trinkt (686), Tee (640), Wein (528), tranken (420), Kaffee (252)
• Uhr	Stunden (19481), Jahre (6444), Tore (4020), Tagen (2826), Grad (1628), Hektoliter (1728), Sekunden (1560), ...
• Umweltschutz	Beitrag (780)
• Verteidiger	Anwalt (3784), Professor (3164), Geschäftsführer (2415), Manager (1590), Pressesprecher (1080)
• Weltrekord	Rekord (460)
• zulässig	bereit (2461), möglich (2112), unzulässig (594), unwirksam (532), verbindlich (182)

Tabelle 3.9: Kollokationen nach dem Signifikanzmaß $\sigma_{2links} \cdot \sigma_{2rechts}$

Durch Anwendung und Kombination der oben eingeführten Signifikanzmaße für Kollokationen zweiter Ordnung kann man eine differenzierte Clusterung der Wörter erreichen, die deren teilweise nicht explizierbare Verbindungen zueinander ausdrückt. Einige interessante Resultate sind:

- „*Duma*“ $\rightarrow \sigma_{2links} \cdot \sigma_{2rechts} \rightarrow$ „*Parlament*“
- „*Uhr*“ $\rightarrow \sigma_{2links} \rightarrow$ Einheiten
- „*Uhr*“ $\rightarrow \sigma_{2rechts} \rightarrow$ Veranstaltungsorte
- „*DM*“ $\rightarrow \sigma_{2links} \rightarrow$ Währungen

3.2.2. Winkel zwischen Kollokationsvektoren

Anstatt des Kreuzproduktes ist es sinnvoll, den Winkel zwischen den Kollokationsvektoren zu berechnen, wenn man Synonyme oder Worte sucht, die einem ähnlichen Kontext verwendet werden:

$$\alpha_{prod}(a,b) = \arccos\left(\frac{\sum_i (\sigma_{CBA}(a,i)\cdot\sigma_{CBA}(b,i))}{|\vec{k}(a)|\cdot|\vec{k}(b)|}\right) \tag{3.13}$$

$$\alpha_{min}(a,b) = \arccos\left(\sum_i min\left(\frac{\sigma_{CBA}(a,i)}{|\vec{k}(a)|},\frac{\sigma_{CBA}(b,i)}{|\vec{k}(b)|}\right)\right) \tag{3.14}$$

Dabei bezeichnet $|\vec{k}(a)|$ die Länge des Kollokationsvektors von a:

$$|\vec{k}(a)| = \sqrt{\sum_i (\sigma_{CBA}(a,i)^2)} \tag{3.15}$$

Resultate:
Dadurch erhält man beispielsweise die Paare *(Riis – Rijs), (Erdbeben – Erdbebenkatastrophe), (spielt – zugedacht),* aber leider auch *(FC – Wissenschaftsstandort)*, da beide starke Kollokationen zu *Bayern* sind oder *(Virchow – Prahm)*, da sie den gleichen Vornamen haben *(Rudolf)*.

3.3. Extraktion semantischer Netze/Cluster aus stark zusammenhängenden Graphen

Die beschriebenen Relationen werden bisher fast ausschließlich als Tabelle der Wörter und dem Signifikanzwert der jeweiligen Relation dargestellt.

Aus der Kollokationstabelle können wir aber auch Graphen um die betrachteten Wörter ableiten. Dazu wurden diejenigen Kollokationen ausgewählt, die mit dem Ausgangswort weitere Kollokationen gemeinsam haben. So werden für einen Graphen um ein Wort a Tripel von Wörtern a, b und c gesucht, bei denen alle drei Wörter zueinander paarweise signifikant auftreten.

Im Programm `create_word_fig` werden mit folgendem SQL-Befehl die Tripel aus der Datenbank geladen:

```
1    select w_a.wort_nr, w_b.wort_nr, w_c.wort_nr,
2       w_a.wort_bin, w_b.wort_bin, w_c.wort_bin,
3       k1.signifikanz, k2.signifikanz, k3.signifikanz,
```

```
4    from wortliste w_a, wortliste w_b, wortliste w_c,
5       kollok_sig k1, kollok_sig k2, kollok_sig k3
6    where w_a.wort_bin='a'
7       and k1.wort_nr1=w_a.wort_nr and k2.wort_nr1=w_a.wort_nr
8       and k1.wort_nr2<k2.wort_nr2
9       and k1.wort_nr2=k3.wort_nr1 and k2.wort_nr2=k3.wort_nr2
10            and k1.wort_nr2=w_b.wort_nr and k2.wort_nr2=w_c.wort_nr
```

In der Tabelle `wortliste` stehen die Wörter in der Spalte `wort_bin`, `wort_nr` bezeichnet den Primärschlüssel der Tabelle. In `kollok_sig` sind die signifikanten Kollokationen auf Satzbasis gespeichert. Einem Paar (`wort_nr1`, `wort_nr2`) ist jeweils ein Signifikanzwert `signifikanz` zugeordnet, der, wie in Abschnitt 3.1.2 auf Seite 25 erläutert, nach dem *Common-Birthday-Maß* berechnet wurde.

Die Tabelle `wortliste` wird für die Wörter a, b und c jeweils als `w_a`, `w_b` und `w_c` referenziert. In der sechsten Zeile wird das Wort a ausgewählt. Je nach Art des Programmaufrufs kann diese Auswahl analog über die Wortnummer erfolgen. In der siebenten Zeile wird `w_a` mit der Kollokationstabelle verknüpft, zum einen unter dem Alias `k1` für das Paar (a, b), zum anderen mit `k2` für (a, c). Die Bedingung in der achten Zeile, dass die Wortnummer von b stets kleiner ist als die von c, verhindert doppelte Tripel in der Ergebnismenge, bei denen lediglich b und c vertauscht sind (das Wort a ist fest gewählt). Als weitere Bedingung wird in der neunten Zeile angefügt, dass in `k3` ein Paar (b, c) existiert. In der letzten Zeile werden nur noch die Wortlisten für b und c mit dem Ergebnis verbunden, um auf die Wörter selbst (und nicht nur auf die Wortnummern) zugreifen zu können.

Der SQL-Befehl führt durch die Bedingung in Zeile 9 zum Ausschluss von singulären Paaren (*Hund – Pawlowschen*), zu denen sich kein drittes Wort finden lässt, das zu beiden eine Kollokation ist. Sie werden ausgeschlossen, da uns in der graphischen Darstellung nicht primär die Stärke der einzelnen Kollokationspaare interessiert, sondern die Struktur der Kollokationenumgebung. So beobachten wir zum Beispiel eine Clusterung in mehrere nur durch das Ausgangswort zusammenhängende Teilgraphen, wenn das Wort in verschiedenen Kontexten (z. B. Eigenname verschiedener Personen oder sowohl Eigenname als auch Berufsbezeichnung) oder mit unterschiedlichen Bedeutungen (Homonym) auftritt.

Anschließend wählt das Programm aus der Menge der vom SQL-Befehl zurückgegebenen Tripel die Kollokationen mit den höchsten Signifikanzen $CB(a, b)$ und $CB(a, c)$ aus. Das Kriterium für diese Auswahl ist die Anzahl der gefundenen Tripel: die Signifikanzen der Kanten vom Ausgangswort müssen größer als der Logarithmus der Anzahl der gefundenen Tripel sein:
$\sigma(a, b) > \log_{10}(\text{Anzahl der Tripel}(a, b, c))$.

Das Paar (b, c) wird nur gespeichert, wenn sowohl das Paar (a, b) als auch (a, c) das

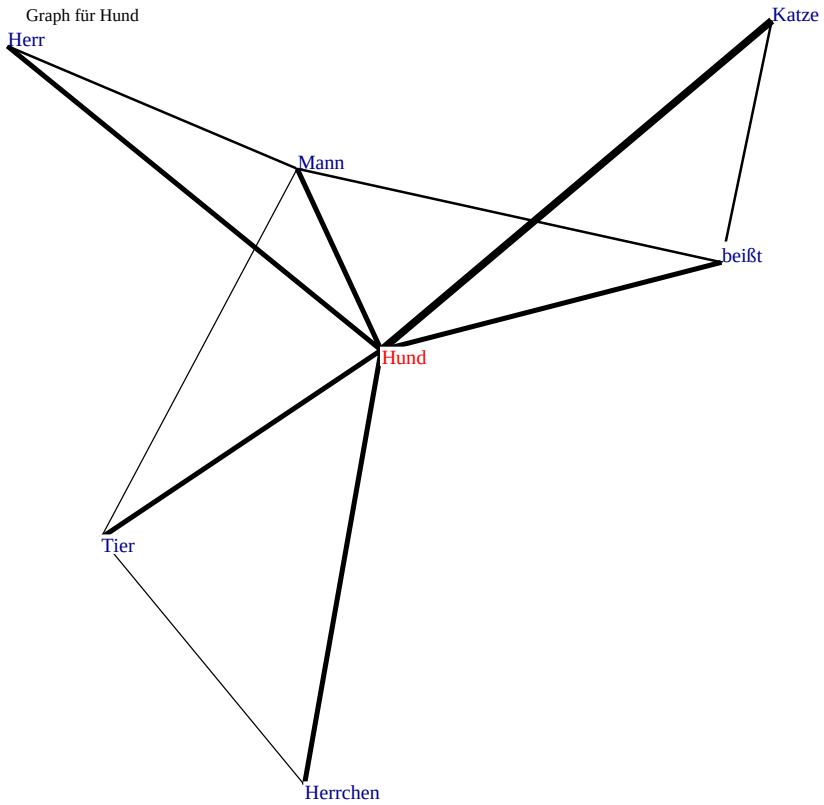


Abbildung 3.1.: Durch simulated annealing erzeugter Graph für *Hund*

Kriterium erfüllen. Bei der Speicherung der Paare werden die doppelte Wörter und Paare ignoriert.

Dieser Schritt führt zur Bevorzugung von Tripeln (*Hund, Mann, beißt*) gegenüber schwächer signifikanten wie (*Hund, Wallace, Gromit*). Dabei ist es nicht bedeutsam, dass das Paar (*Wallace – Gromit*) einen sehr hohen Kollokationswert hat.

Der Ausschluss schwacher Verbindungen kann dazu führen, dass Wörter dargestellt werden, die scheinbar nur Kollokationen zum Ausgangswort sind und mit keinem der anderen dargestellten Wörter in Relation stehen. Diese Wörter werden als stärksten Vertreter ihres Teilgraphen trotzdem dargestellt (siehe Abbildung 3.2). Der Teilgraph kann aber nicht komplett dargestellt werden, weil der Gesamtgraph sonst zu komplex und damit zu unübersichtlich würde.

Auf die Positionierung der Wörter in der graphischen Darstellung wird im Abschnitt 4.2 eingegangen, der erläutert, wie mit der Methode des *Simulated annealing* aus den Informationen über die Wörter und der Stärke ihrer Verbindungen eine Position des Wortes im zweidimensionalen Raum errechnet wird.

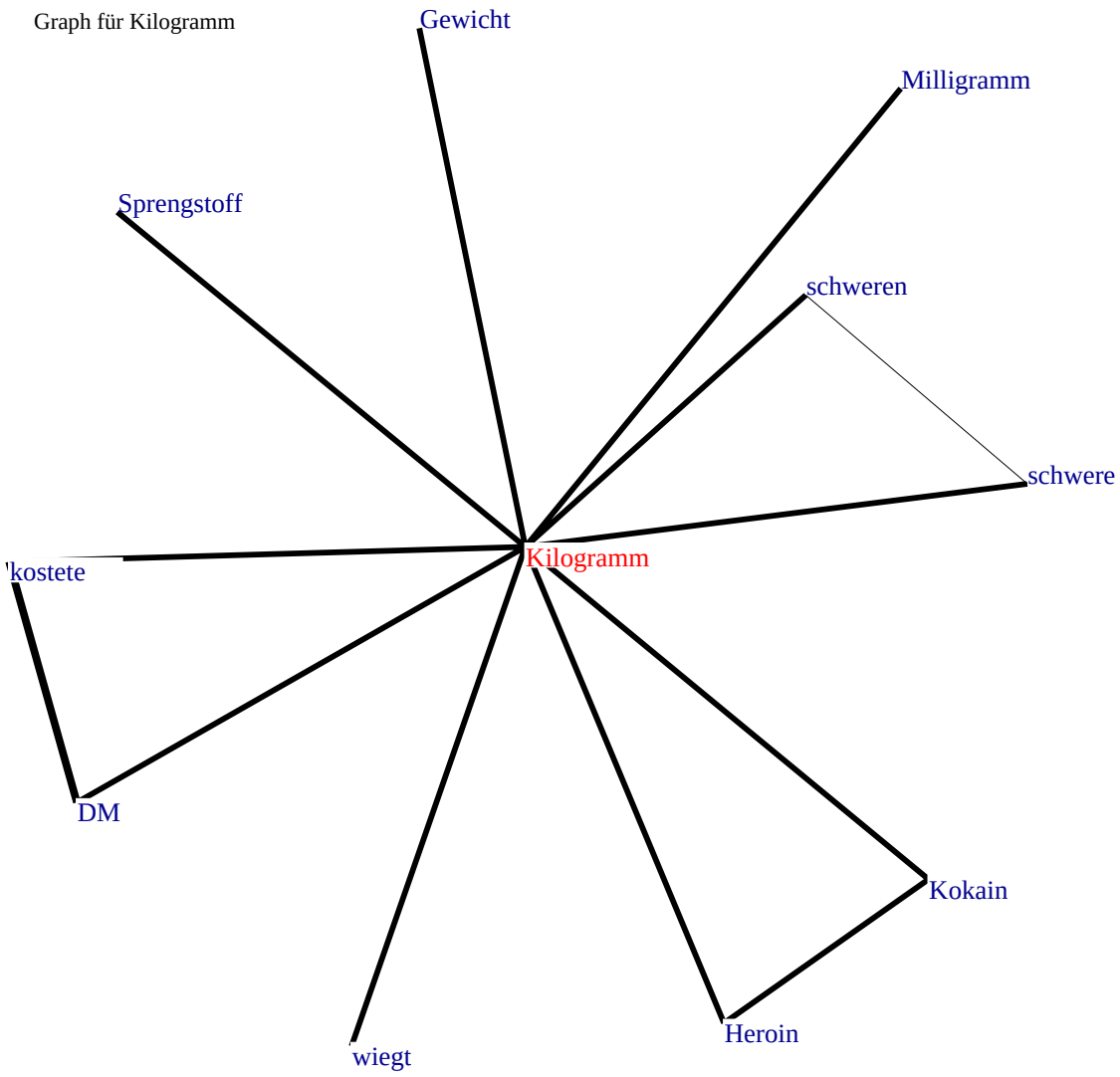


Abbildung 3.2.: Durch simulated annealing erzeugter Graph für *Kilogramm*

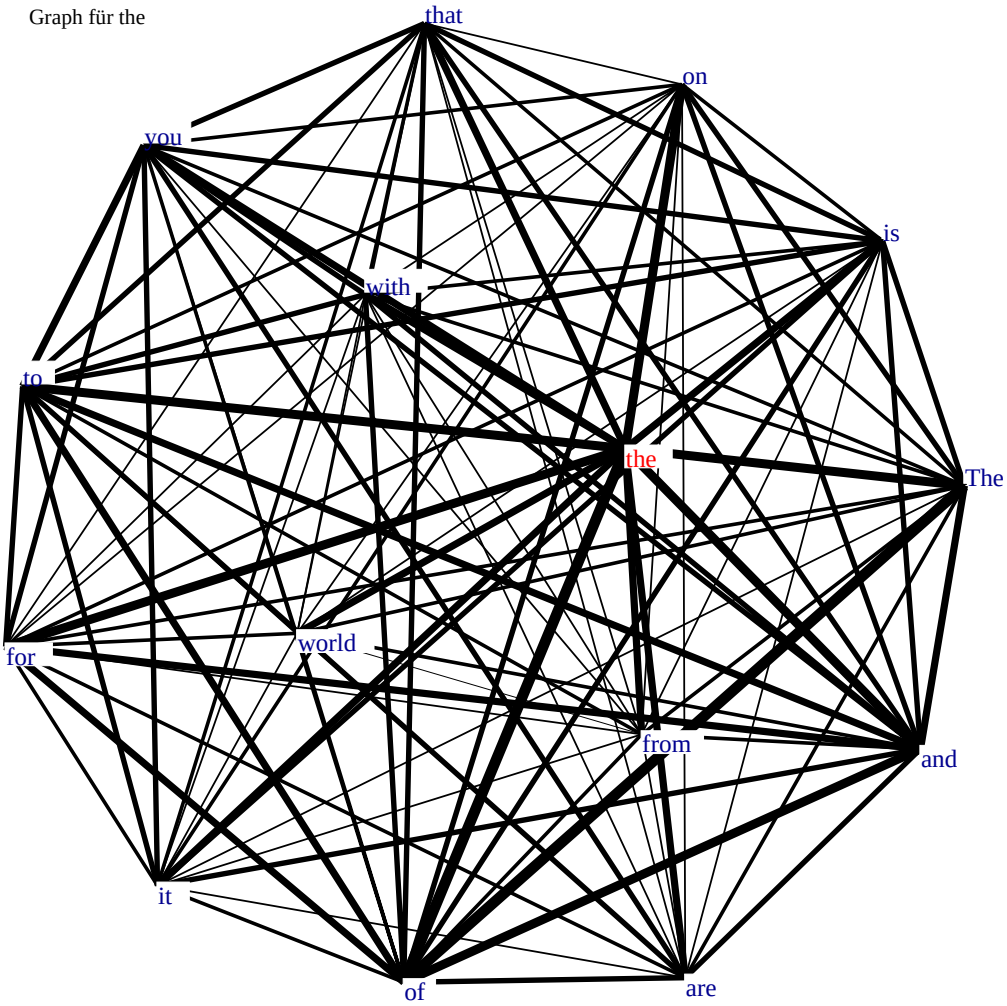


Abbildung 3.3.: Durch simulated annealing erzeugter Graph für *the*

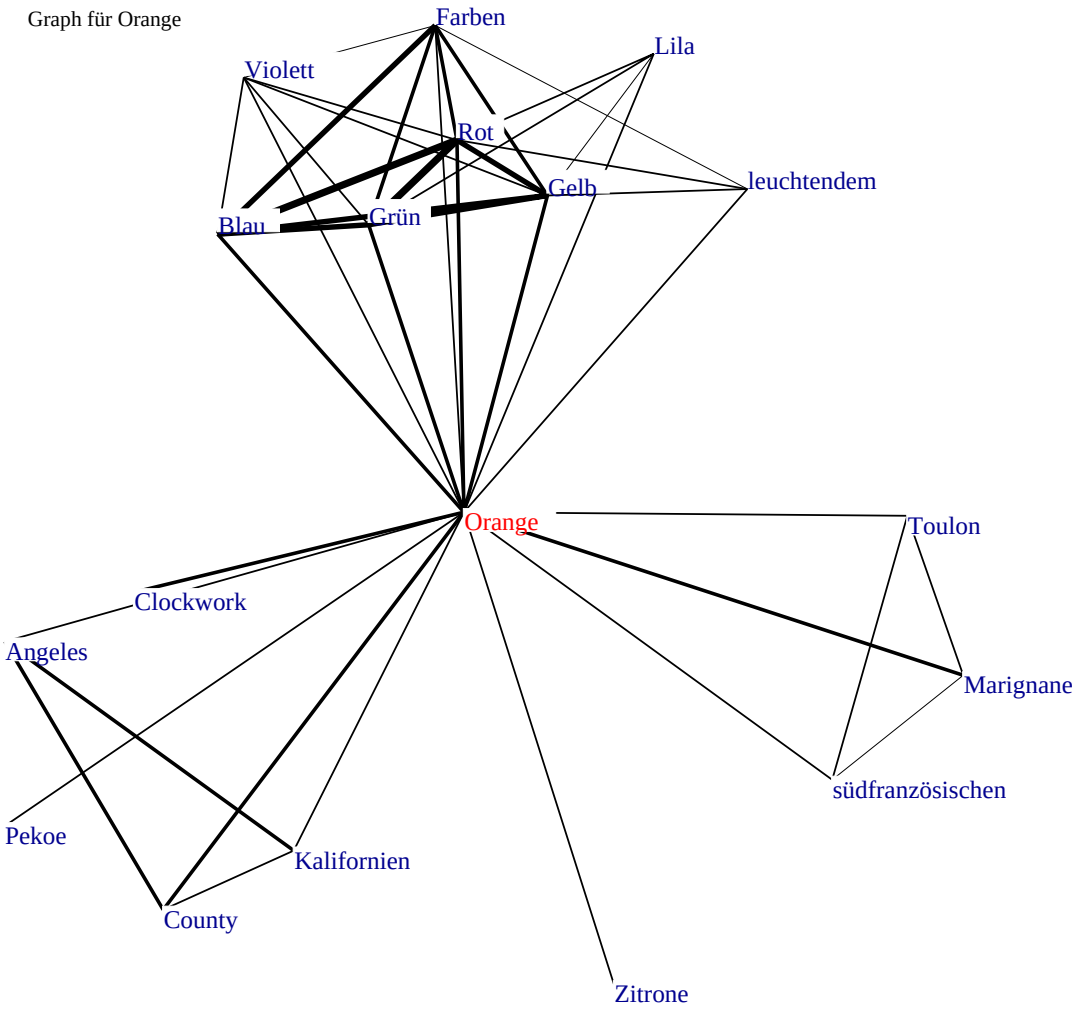


Abbildung 3.4.: Durch simulated annealing erzeugter Graph für *Orange*

3.3.1. Cluster

Eine weiterführende Stufe zur Darstellung von Kollokationsgraphen ist die Suche nach Clustern signifikanter Kollokationen, in denen auch Wörter enthalten sind, die nicht direkt mit dem Ausgangswort in Beziehung stehen.

Dazu muss eine Bewertungsfunktion aufgestellt werden, mit deren Hilfe man entscheiden kann, welche Kollokationen zu einem Cluster hinzugefügt werden. Diese muss den Zusammenhang mit (prozentual) vielen Komponenten des Clusters gegenüber der Stärke der Verbindung zwischen den Paaren sehr hoch bewerten. Im Rahmen dieser Arbeit wurden diesbezüglich keine Untersuchungen vorgenommen.

Durch die Fülle der vorhandenen Kollokationspaare und der Notwendigkeit, auch weniger signifikante Paare zu untersuchen, ergibt sich eine hohe Komplexität im Auffinden solcher Cluster.

3.4. Exkurs: Kombination der statistischen Methoden mit explizitem Wissen

Die statistischen Signifikanzmaße in Abschnitt 3.1 beschränkten sich auf die Extraktion relevanter Kollokationen, die räumlich benachbart im Textkorpus stehen. In Abschnitt 3.2 wurde bei der Berechnung semantischer Relationen die Beschränkung auf räumliche Nachbarschaft überwunden. So konnten Wörter gefunden werden, die in ähnlichem Kontext stehen oder die ähnlich gebraucht werden. So wurden zu *Minister* nicht nur der *Ministerpräsident* und die Bezeichnung von Ministern bestimmter Ministerien (wie z. B. *Landwirtschaftsminister*) gefunden, sondern auch der *Trainer*.

Nun sind die Aufgben eines Trainers zwar mit denen eines Ministers vergleichbar. Beispielsweise *sagen, meinen* oder *erklären* beide oft irgend etwas, bei Trainer wie bei Ministern interessieren die Öffentlichkeit weiterhin *frühere* oder *ehemalige* ebenso wie potentielle *Nachfolger*. Diese Beispiele lassen sich noch durch 19 andere gemeinsame Kollokationen ergänzen. Trotzdem gehört der *Trainer* oft nicht zu den erwünschten Antworten zu *Minister*.

Eine Möglichkeit, diese Resultate der Kollokationen zweiter Ordnung weiter zu bewerten, ist die Anwendung expliziten Wissens. Im Wortschatz-Lexikon sind zu vielen Wörtern Sachgebietsangaben gespeichert. Mit deren Hilfe können bei der Rückgabe der gesuchten Wörter die herausgefiltert werden, die nicht im gleichen Sachgebiet wie das Ausgangswort stehen.

Obwohl zu vielen Wörtern eine Sachgebietsangabe existiert, sind diese Sachgebiete zum Teil so speziell, dass nur wenige andere Wörter zum gleichen Sachgebiet gehören. So kennen wir aus dem Sachgebiet *Endokrinologie* nur 14 Wörter, zu de-

nen bei keinem das Sachgebiet *Medizin* vorliegt. Deshalb würden fast alle Wörter herausgefilter werden, wenn man das gleiche Sachgebiet fordert. Um festzustellen, ob Wörter zu ähnlichen Gebieten gehören, haben wir die Sachgebiete in einer Hierarchie angeordnet. So konnten wir einen Ähnlichkeitsgrad einführen, indem wir die Anzahl der übereinstimmenden Hierarchieebenen zählten, in denen die Sachgebiete stehen. Wenn zu einem Wort weitere Sachgebiete angegeben sind, wird die Anzahl weiterer übereinstimmender Hierarchieebenen addiert.

Beispiel:
Das Wort *Molekularbiologie* gehört zu den Gebieten *Biologie, Medizin* und *Biochemie/Biophysik/Cytologie*, das Wort *Gen* zu *Vermessungswesen, Medizin, Biochemie/Biophysik/Cytologie, Genetik/Evolution* und *Biologe*. Explizit haben die Wörter zwei gemeinsame Sachgebiete: *Medizin* und *Biochemie/Biophysik/Cytologie*. Diese gehören zu den Gebieten *Biologie*[2], *Naturwissenschaft* und *Wissenschaft*. Damit ergibt sich die Anzahl von fünf gemeinsamen Sachgebieten.

Dieses Ähnlichkeitsmaß ermöglicht es, die Kollokationen (oder Wörter aus anderen Relationen) umzuordnen oder Wörter ohne gemeinsames Sachgebiet herauszufiltern. Vorher ist abzuschätzen, ob die Datenbasis der Sachgebietsangaben ausreicht, um auch die gewünschten Ergebnisse zu erhalten. Da in der Regel nicht für alle Wörter Sachgebiete vorliegen, nimmt man mit der Verbesserung des gefundenen Wörter auch einen Verlust möglicher richtiger Wörter in Kauf (das *Precision-Recall-Verhältnis* ändert sich).

Da wir bei der Ermittlung der Kollokationen mit den Vollformen der Wörter arbeiten, die Sachgebiete aber nur für Grundformen vorliegen, wird dieser Ansatz im Wortschatz-Projekt noch nicht verwendet. Neben der Grundformreduktion ist auch eine noch größere Datenbasis an Sachgebietsangaben nötig, damit nicht zu viele Sachgebiete ausgefiltert werden.

Neben Sachgebietsangaben kann auch anderes Wissen auf analoge Art verwendet werden. So können Relationsbäume aufgebaut werden, die nicht nur Sachgebietsangaben enthalten, sondern weitere, in der Wissensrepräsentation übliche Relationen (`is-a`, `part-of` etc.) oder eine Ontologie verwendet werden. Bei der Berechnung des Ähnlichkeitsmaßes sollte hierbei aber beachtet werden, welche Relationen transitiv sind, also über welche Hierarchieebenen die Anzahl höherer Ebenen addiert werden kann.

Eine weitere Verfeinerung ist eine Gewichtung der Relationen, also der Kanten im Baum. Dann wird nicht mehr die Anzahl identischer Hierarchieebenen sondern die Werte der Relationen aufsummiert. Untersuchungen zu den Möglichkeiten des Aufbaus einer Ontologie im Rahmen des Wortschatz-Projektes stehen noch aus.

[2] *Gen* ist auf Grund eines Schreibfehlers nur im Sachgebiet *Biologe*, nicht aber in *Biologie*

4. Darstellungsverfahren

4.1. Darstellung geradliniger, ungerichter Graphen

Um die in Abschnitt 3.3 abgeleiteten semantischen Netze darzustellen, suchten wir nach einem effizienten Algorithmus, der ästhetische Graphen erzeugt. Zunächst soll der Begriff der *Ästhetik* eines Graphen präzisiert werden, indem einige objektive Beurteilungskriterien oder *Ästhetiken* nach [CT94, S. 10 ff.] aufgezeigt werden.

- niedrige Anzahl der Kreuzungspunkte der Kanten
- geringe Ausdehnung des Gesamtgraphen
- Maximierung des kleinsten Winkels zwischen benachbarten Kanten des Graphen
- symmetrische Auszulegung symmetrischer Teilgraphen
- stark zusammenhängende Knoten liegen nah beieinander

Wie diese sich oft widersprechenden Kriterien im einzelnen angewendet und wie sie gegeneinander gewichtet werden, hängt von der Anwendung und der gewählten Darstellungsart ab. So haben beide Zeichnungen in Abbildung 4.1 ihre Berechtigung, je nachdem ob die Relationen zu einem gegebenen Wort (das bei der Betrachtung „im

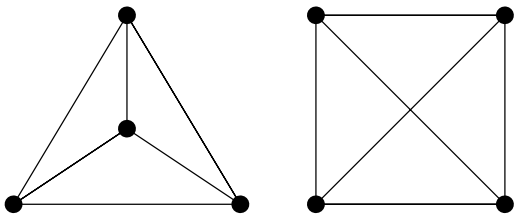


Abbildung 4.1.: kongruente Graphen, optimiert nach minimaler Kreuzungszahl und symmetrischer Darstellung

Zentrum steht“) oder die Relationen zwischen Wörtern einer Wortmenge dargestellt werden oder „im Vordergrund stehen“.

In Abbildung 4.2 wird in vielen Anwendungsfällen die rechte Darstellung bevorzugt werden, die den Graphen als dreidimensionale Struktur darstellt, auch wenn hier weder Kreuzungspunkte vermieden noch eine symmetrische Darstellung gewählt wurde. Die dreidimensionale Wahrnehmung einer zweidimensionalen Abbildung erscheint zwar als schwierigere Aufgabe, aber nach der Wahrnehmungspsychologie wird ein Objekt immer auf die einfachste mögliche Weise wahrgenommen. Ein Würfel ist eine einfachere Repräsentation der rechten Figur in Abbildung 4.2 als beispielsweise die Repräsentation als eine Menge von einem Quadrat, zwei Dreiecken und vier Trapezen.

Wenn ein beliebiger Graph dargestellt werden soll, ist es aber schwierig, a priori ein Kriterium dafür zu finden, ob eine zwei- oder dreidimensionale Strukturen diesen Graphen ästhetischer und leichter wahrnehmbar repräsentiert.

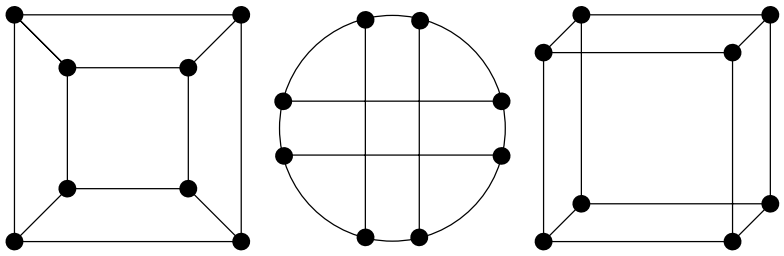


Abbildung 4.2.: Darstellungen eines $2 \times 2 \times 2$–Würfels

4.2. Die Simulated-Annealing-Methode zur Erzeugung von Graphen

Zum Zeichnen der Graphen wird im Projekt *Deutscher Wortschatz* die Methode des *simulated annealing* bzw. das *force directed placement* (Bezeichnung der Variante des Simulated-Annealing-Ansatzes von Fruchtermann & Reingold) verwendet. Diese beruht aus der Optimierung eines Kräftegleichgewichts zwischen den Knoten des Graphen (in diesem Fall zwischen den Positionen der Wörter). Man kann den Graphen mit einem Atomgitter vergleichen, in dem Atome auf Grund ihrer gleichen elektrischen Ladung voneinander abgestoßen werden, benachbarte Atome aber durch gemeinsame Elektronen zusammengehalten werden. Die Anziehungskräfte herrschen zusätzlich zu den Abstoßungskräften.

Im ursprünglichen Modell von Eades [Ea84] werden benachbarte Knoten mit Federn einer bestimmten Länge verbunden, die die optimale Entfernung zwischen den

Knoten beschreiben; nicht benachbarte Knoten sind mit Federn unendlicher Länge verbunden. Um die optimale Position der Knoten zu finden, werden die aus dem Modell resultierenden Differentialgleichungen gelöst oder das System „entwickelt“ bzw. simuliert.

Bei der Positionsbestimmung mittels *simulated annealing* wird ein zusätzlicher Temperaturfaktor eingeführt, der im Verlauf der Simulation abgekühlt wird (engl.: annealing = Ausglühen). Zu Beginn der Simulation herrscht im Atomgitter eine hohe Temperatur, so dass die Knoten stark um ihren Ausgangspunkt schwingen. In Abhängigkeit der wirkenden Kräfte und der Temperatur des Systems werden die Knoten in jedem Iterationsschritt zur optimalen Posisition hinbewegt. Durch eine anfangs sehr hohe Temperatur soll verhindert werden, dass die Knoten sich in einem lokalen Minimum stabilisieren, indem der Knoten durch hohe Abstoßungskräfte anderer Knoten daran gehindert wird, sich seinen Nachbarn zu nähern (d. h. Schlaufen im Graphen entwirren sich nicht). Wenn die Temperatur weiter abgekühlt ist, bleibt die relative Lage der Knoten zueinander stabil, und es werden nur noch die Abstände optimiert.

Dieses Verfahren lässt sich durch folgenden Algorithmus formulieren:

Solange der Graph nicht abgekühlt ist:
berechne für jeden Knoten aus der Entfernung zu den anderen Knoten die Größe und Richtung der an diesem Knoten wirkenden abstoßenden Kräfte
berechne für jeden Knoten aus der Entfernung aller benachbarten Knoten die anziehenden Kräfte
verschiebe gleichzeitig alle Knoten in Abhängigkeit der wirkenden Kräfte und der Temperatur
kühle die Temperatur ab

Die Temperatur T des Graphen stellt in Abhängigkeit von der Anzahl der bisherigen Iterationen t eine sigmoide Funktion dar. Diese ist linear abhängig von der Anzahl der darzustellenden Knoten a (Wörter):

$$T = \frac{a}{2 \cdot (1 + e^{t/8-5})} + T_0 \tag{4.1}$$

T_0 bezeichnet die minimale Temperatur des Graphen.

Abbildung 4.3 stellt einen typischen Verlauf der Temperaturkurve dar, hier für einen Graphen mit 25 Knoten.

Eades verwendet zur Berechnung der anziehenden Kräfte F_1 und der abstoßenden Kräfte F_2 folgende Formeln:

$$F_1(d) = k_1 \log(d/k_2) \tag{4.2}$$
$$F_2(d) = k_3/d^2 \tag{4.3}$$

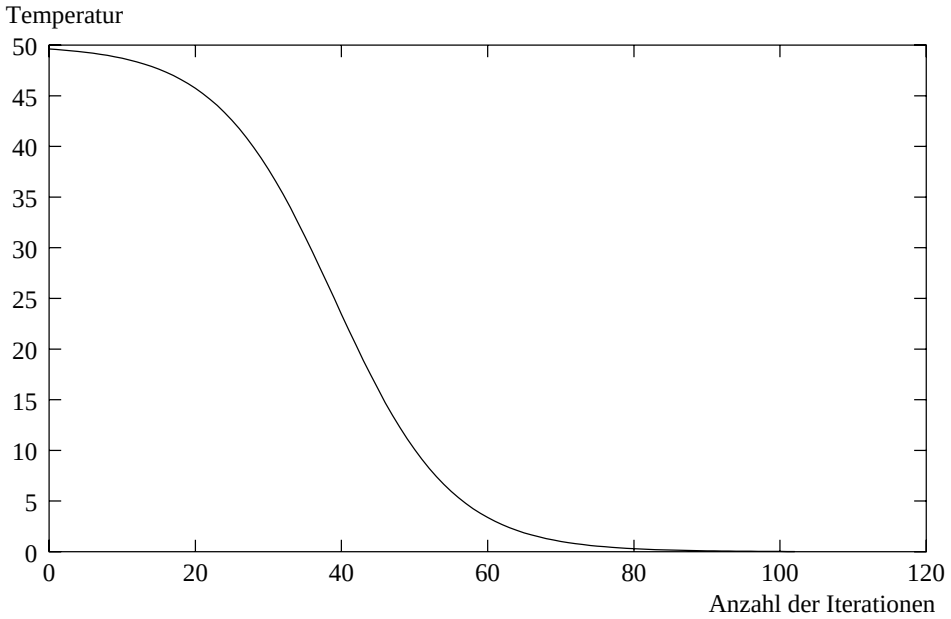


Abbildung 4.3.: Abhängigkeit der Temperatur des Graphen von der Anzahl der Iterationen

d bezeichnet in die alten Abstände zweier Knoten, k_i sind Konstanten, mit denen das Verfahren den eigenen Bedürfnissen angepasst werden kann. Die abstoßenden Kräfte berechnet Eades nur für Knoten, die nicht verbunden sind. Fruchtermann & Reingold verwenden:

$$F_1(d) = d^2/k \tag{4.4}$$
$$F_2(d) = -k^2/d \tag{4.5}$$

Wieder bezeichnen d den Abstand zweier Knoten, F_1 die anzeihenden Kräfte und F_2 die abstoßenden Kräfte, die in diesem Ansatz jedoch werden auch zwischen benachbarten Knoten berechnet werden. Mit k wird der optimale Abstand zweier Knoten im Graphen gewählt, der aus der Anzahl der Knoten und der Größe der fertigen Zeichnung berechnet. In die Berechnung von k sollte eingehen, wie sich der Graph entfalten kann, d. h. ob viele der Knoten benachbart sind und der Graph deswegen ein Knäuel verbundener Knoten bleibt.

Das Verfahren kann keinen optimalen Graphen garantieren, aber erzeugt mit einem geringen Rechenaufwand einen ästhetisch akeptablen, insbesondere kreuzungsarmen Graphen. Eine Möglichkeit der Optimierung besteht in der expliziten Berücksichtigung des Winkels zwischen zwei von einem Knoten ausgehenden Kanten wie bei Davidson & Harel [DH96]. Den Algorithmus kann man auch für eine dreidimensionale Darstellung der Graphen anwenden.

Eine Vorstellung des ursprünglichen Verfahrens von Eades und eine Diskussion der

verbesserten Varianten von Fruchterman & Reingold und Kamada & Kawai nebst der Vorstellung eines eigenen Ansatzes für eine Optimierung des Verfahrens von Eades findet sich bei Davidson & Harel [DH96].

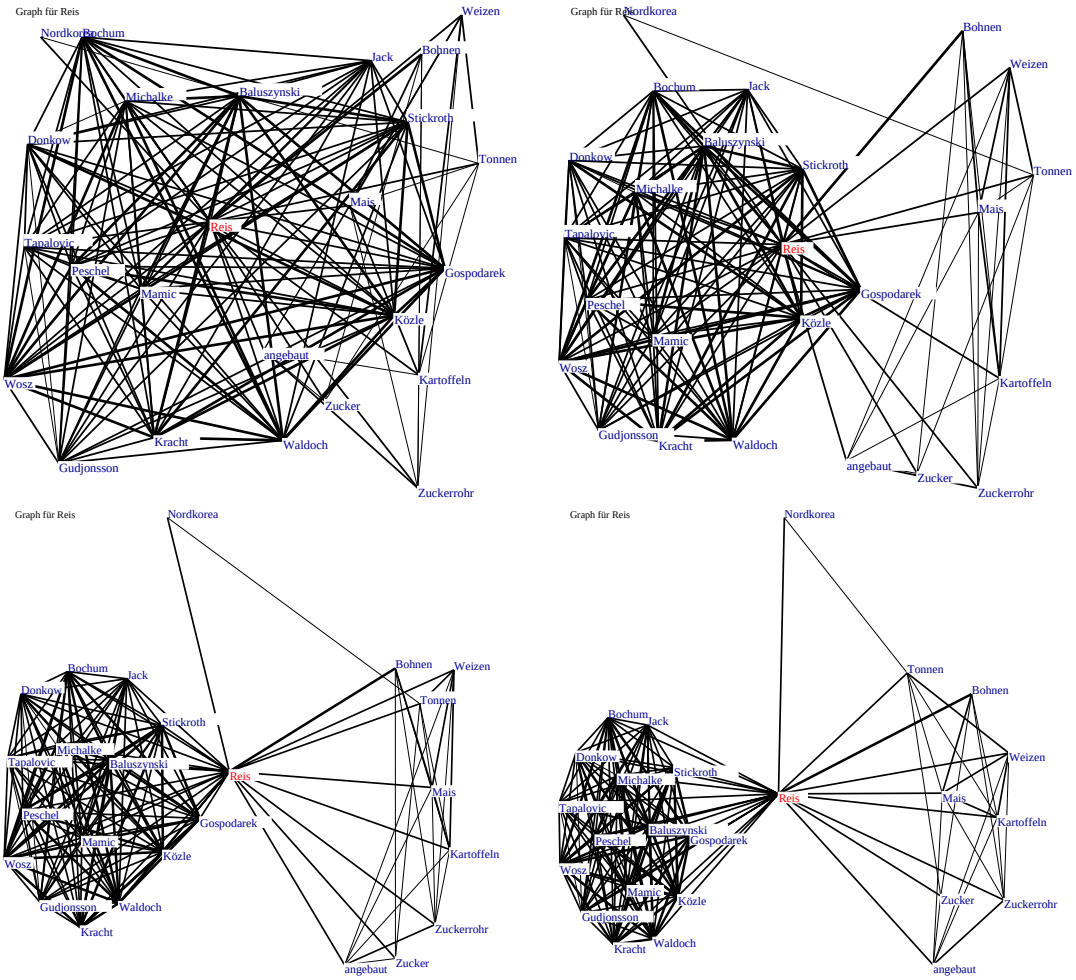


Abbildung 4.4.: Entstehungsfolge des Graphen für *Reis*: zufällig initialisierter Graph, Graphen nach 15 und 30 Iterationen, fertiger Graph

4.3. Erzeugung des Kollokationsgraphen

Im Abschnitt 3.3 auf Seite 41 wurde bereits erläutert, wie aus den Kollokationen Tripel zur Darstellung eines Graphen ausgewählt werden. Die Wörter und Kollokationen werden als Knoten und Kanten in Adjazenzlistendarstellung [DuInf88, S. 254] gespeichert. Anschließend wird auf diesen Graphen der Simulated-Annealing-Algorithmus (Abschnitt 4.2) solange angewendet, bis die Temperatur des Graphen unter 0,02 gesunken ist.

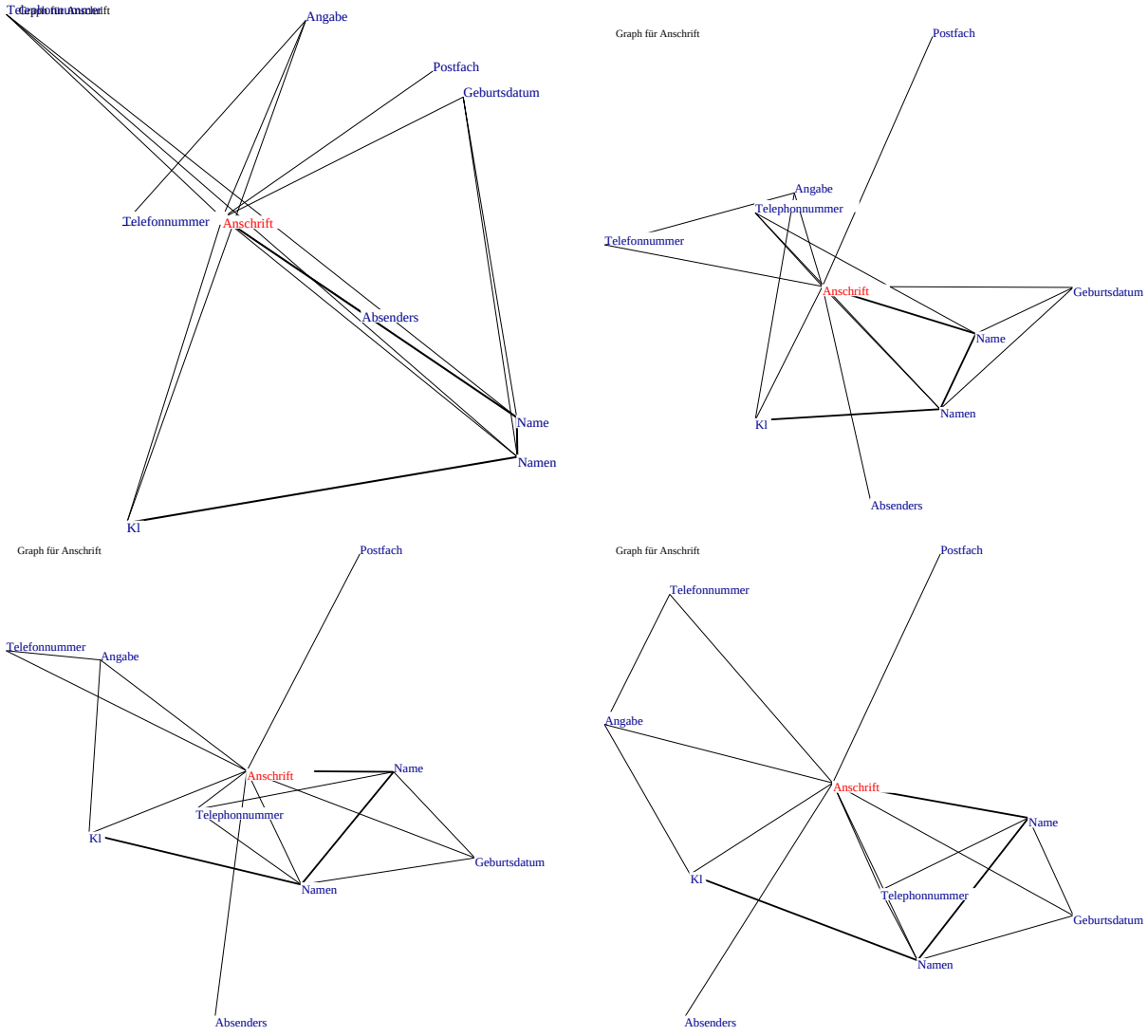


Abbildung 4.5.: Entstehungsfolge des Graphen für *Anschrift*: zufällig initialisierter Graph, Graphen nach 15 und 30 Iterationen, fertiger Graph

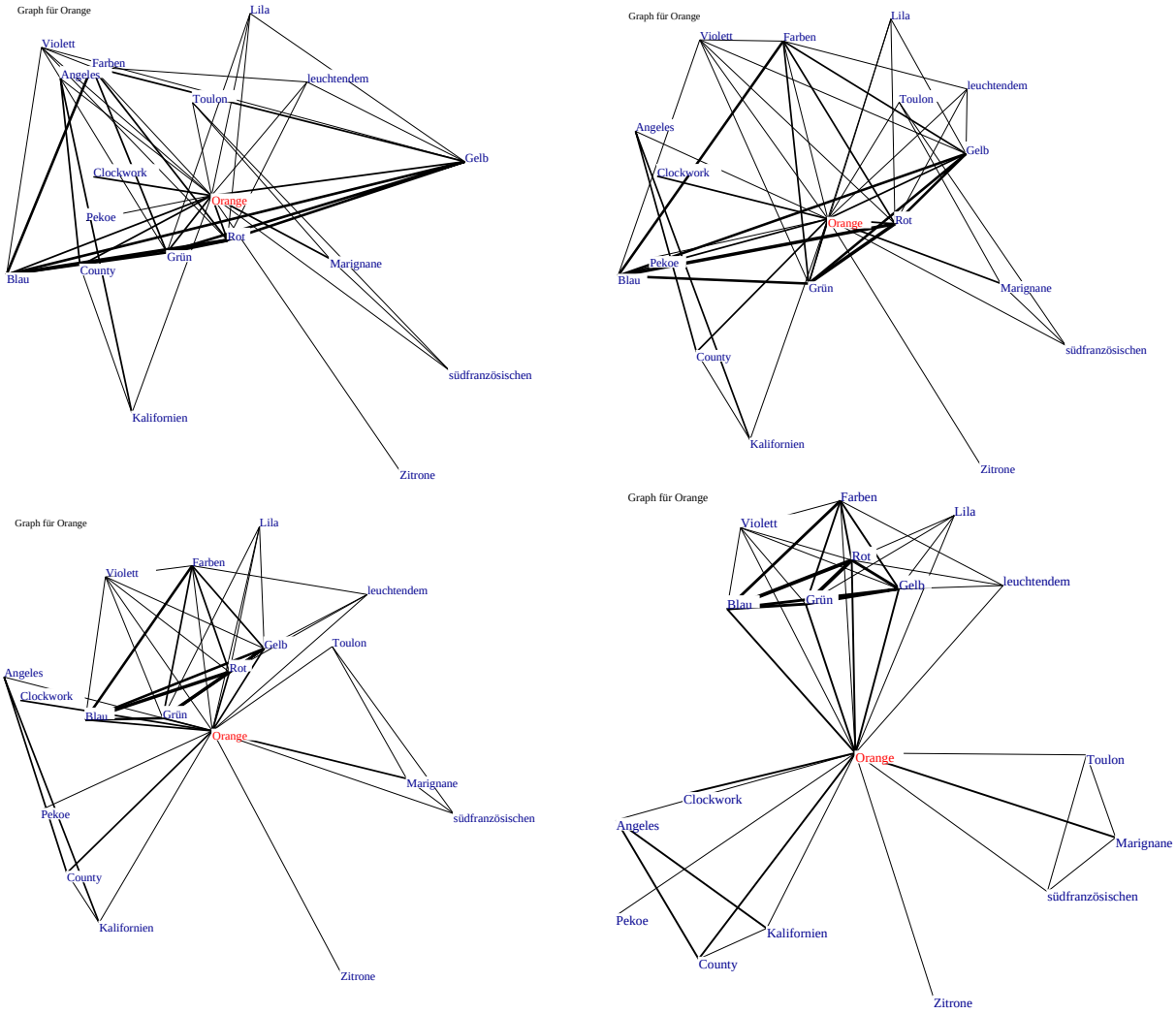


Abbildung 4.6.: Entstehungsfolge des Graphen für *Orange*: zufällig initialisierter Graph, Graphen nach 11 und 21 Iterationen, fertiger Graph

Schließlich werden das Koordinatensystem der Zeichnung so transformiert, dass die Größe des Graphen stets konstant ist und die Knoten und Kanten im `xfig`-Format[1] ausgegeben. In der fertigen Darstellung des Netzes wird das Ausgangswort farblich hervorgehoben, während die Breite der Verbindungslinien zwischen den Knoten mit der Signifikanz der Kollokationen korrespondiert (Breite $= \log(\sigma)$).

Zur Ausgabe des Graphen werden folgende Zeichnungselemente benutzt:

```
# Kante von einem Punkt (x1, y1) nach (x2, y2):
2 1 0 Breite 0 7 0 0 -1 0.000 0 0 -1 0 0 2
         x1 y1 x2 y2
# Hinterlegung der Beschriftung der Knoten
# mit einer weißen, gefüllten Box:
2 2 0 1 7 7 0 0 20 0.000 0 0 -1 0 0 5
         x1 y1 x2 y1 x2 y2 x1 y2 x1 y1
# Beschriftung der Knoten:
4 0 0 0 0 0 Schriftgrad 0.0000 4 Hoehe Breite x1 y1 Wort\001
```

Das Programm `create_word_fig` kann direkt aufgerufen werden, um den Graphen als `xfig`-Datei zu speichern. Dazu wird die Wortnummer oder das Wort selbst als Argument übergeben. Alternativ kann das Skript `wort_graph.pl` benutzt werden, das `create_word_fig` aufruft, um den Graphen zu erzeugen und ihn anschließend mit `fig2dev` in ein anderes Grafikformat umzuwandeln. Dieses wird als zweites Argument übergeben. Für die WWW-Oberfläche werden die Graphen in `gif`-Dateien konvertiert, für die lokale Betrachtung des Graphen sind ebenso die Formate Postscript, JPEG oder TIFF möglich. `transfig` ist ein Programm von Micah Beck (Cornell University), das `xfig`-Dateien in andere Grafikformate umwandeln und skalieren kann.

4.4. WWW-Interface des Projektes Deutscher Wortschatz

Um die Daten des Wortschatz-Projektes plattformunabhängig und ohne zusätzlichen Installationsaufwand nutzbar zu machen, wurde ein Programm erstellt, mit dem die Daten im World Wide Web abgefragt werden können. Zur Übermittlung der Daten baut das Programm auf dem *Common Gateway Interface (CGI)* auf.

CGI basiert auf einer Vereinbarung der Entwickler von HTTP-Servern. Es ist eine Schnittstelle zwischen Informationsservern und Programmen, um deren Ausgabe

[1] `xfig` ist ein einfaches, objektorientiertes Zeichenprogramm, das auf den meisten UNIX-Varianten verfügbar unter http://www-epb.lbl.gov/xfig/; das Datei-Format der jeweils aktuellen `xfig`-Version ist unter http://www-epb.lbl.gov/xfig/fig-format.html dokumentiert

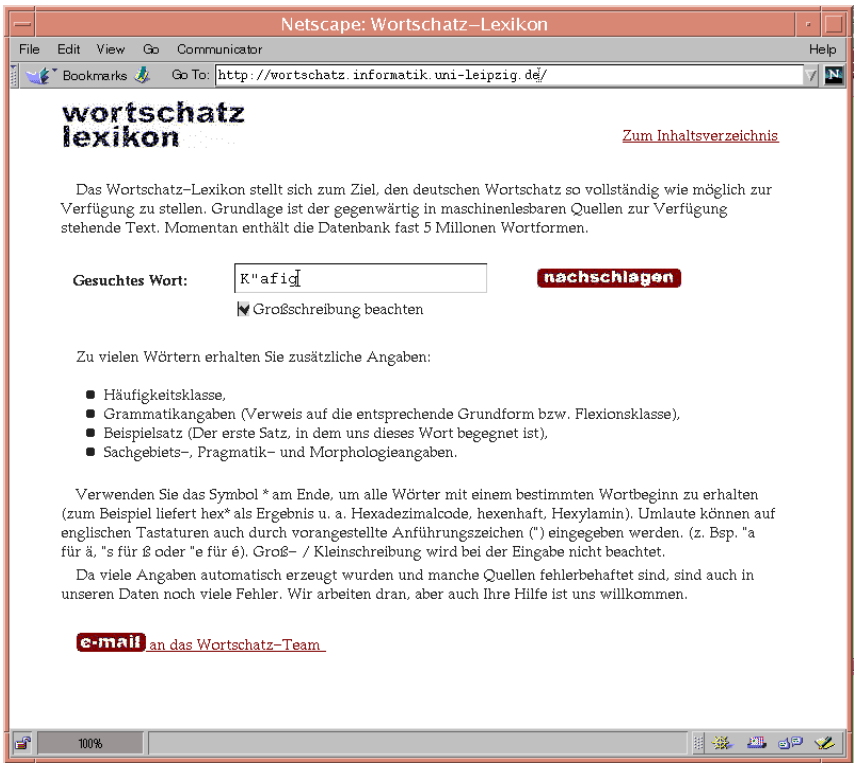


Abbildung 4.7.: Abfragefenster der Wortschatz-Oberfläche

den Nutzern des Informationsanbieters zu präsentieren. Dadurch wurde es möglich, HTML-Dokumente dynamisch zu erzeugen, die vor der Einführung von CGI als statische Dokumente auf dem Server gespeichert werden mussten. Der Standard regelt weiterhin, wie aus einer HTML-Seite Parameter an das Programm übergeben werden können. Er ist unter `http://hoohoo.ncsa.uiuc.edu/cgi/overview.html` spezifiziert.

Zur Präsentation der Wortschatz-Daten wurde ein Programm entworfen, das über die CGI-Schnittstelle durch ein HTML-Formular aufgerufen wird. Es extrahiert die angeforderten Daten aus der Wortschatz-Datenbank und bereitet sie für die Darstellung im WWW-Browser auf. Das Programm wurde in der Sprache C implementiert, da so bei entsprechender Programmierung eine Portierbarkeit auf andere Plattformen möglich ist. Im Gegensatz zu Perl oder Java existieren für C effizientere Compiler.

Wortsuche

Um nach Wörtern zu suchen, steht den Nutzern ein Eingabefeld zur Verfügung. Zum einen kann hier ein einfaches Wort eingegeben werden, zu dem die zugehöri-

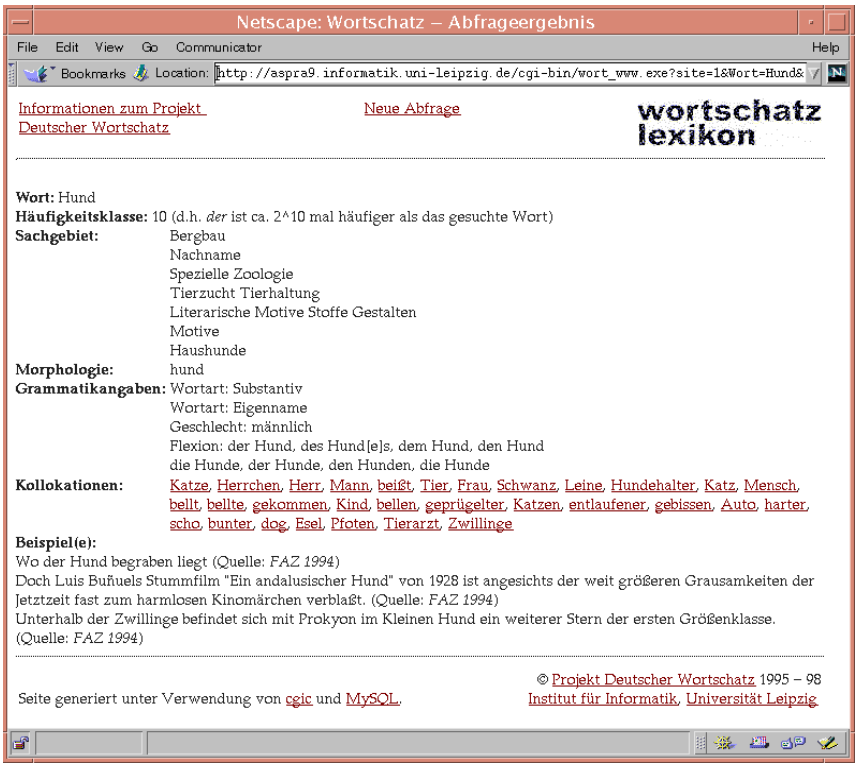


Abbildung 4.8.: Ergebnis für die Abfrage des Wortes *Hund*

gen Datenbankinformationen abgerufen werden sollen. Zur Eingabeerleichterung für Nutzer ohne deutsche Tastatur können die Sonderzeichen `ß`, `ä`, `ö`, `ü` und `é` bzw. `Ä`, `Ö`, `Ü` und `É` durch ein vorangestelltes `"` eingegeben werden, also beispielsweise `gem"a"s` für `gemäß` oder `Caf"e` für `Café`.

Zum anderen kann auch nach Mengen von Wörtern gesucht werden, indem Jokerzeichen als Platzhalter verwendet werden. Dabei repräsentieren `*` oder `%` Zeichenketten mit variabler Länge, während `?` oder `_` anstelle von einzelnen Zeichen eingesetzt werden können. Dies ermöglicht sowohl die Bildung von Ausdrücken, wie sie in der Standardabfragesprache SQL üblich sind, als auch die Formulierung von Ausdrücken, die den regulären Ausdrücken wie sie Betriebssysteme verwenden, ähnlich sind.

Die durch die Abfrage gefundenen Wörter werden als sortierte Liste ausgegeben (siehe Abbildung 4.9) und mit einem Link versehen, der auf die Informationen zu den einzelnen Wörtern zeigt und den Nutzern die nochmalige Eingabe des gesuchten Wortes erspart. Werden mehr als 20 Wörter gefunden, gibt das Programm – wie bei Suchmaschinen üblich – nur die ersten Wörter aus, gefolgt von einem Link auf die Seite mit den nächsten Wörtern.

Falls das gesuchte Wort nicht in der Datenbank enthalten ist, wird der Nutzer ge-

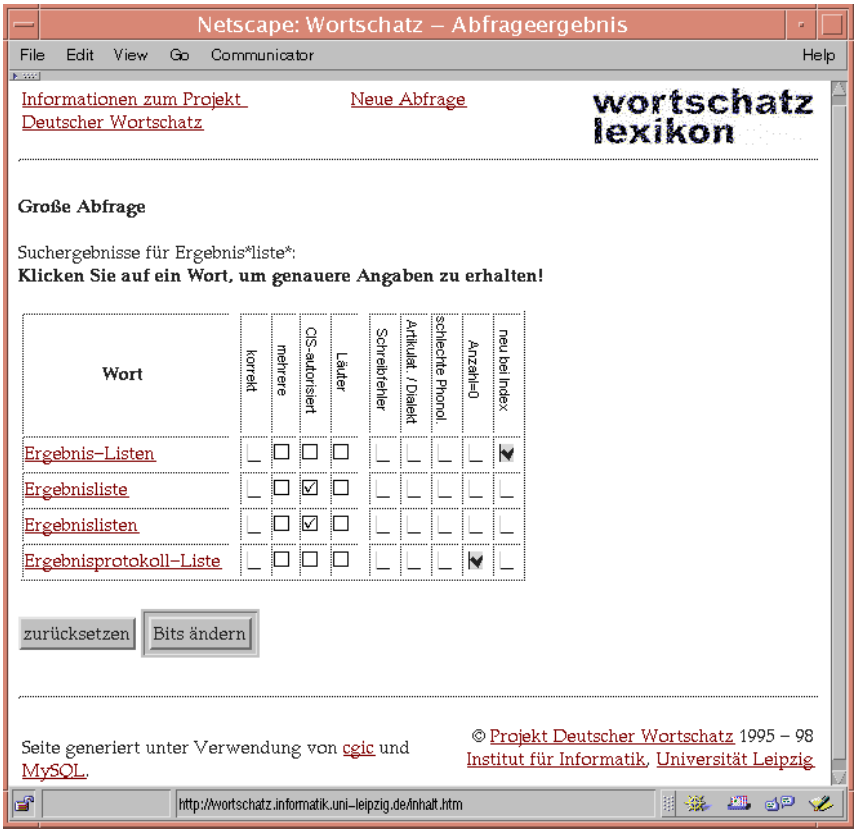


Abbildung 4.9.: Darstellung des Ergebnisses für die Abfrage von *Ergebnis*liste**

beten, das Wort als neues Wort vorzuschlagen. Dazu kann der Nutzer beliebige Angaben zum Wort machen (Beispielsatz, Grammatikangaben etc.), die zusammen mit dem Wort unstrukturiert gespeichert werden. Dieser Eintrag wird erst nach einer redaktionellen Bearbeitung des Wortschatz-Teams in die Datenbank aufgenommen.

Außerdem wird bei nicht vorhandenen Wörtern geprüft, ob eventuell ein ähnliches Wort in der Datenbank enthalten ist. Dazu werden die Wörter der Datenbank in einen Suchbaum geladen. In diesem wird nach Varianten des Ausgangswortes gesucht, die dann dem Nutzer für eine neuen Anfrage vorgeschlagen werden. Filgende Varianten werden dabei berücksichtigt:

- zwei benachbarte, vertauschte Buchstaben (`Wäldre` statt `Wälder`)
- ein eingefügter oder ausgelassener Buchstabe (`Gemeinderatsmittglied`, `Passivlegitmation` oder ähnliches)
- ein vertippter Buchstabe (`narrem` statt `narren`)
- Verwendung von der Schreibweise ohne Sonderzeichen (`ae` statt `"a` etc.) und

Vertauschung von `f` und `ph` (`Filosophie` statt `Philosophie`, `Filosofie` oder `Filzosophie`)

In den folgenden Abschnitten werden die zu einzelnen Wörtern vorhandenen Informationen näher erläutert.

Häufigkeitsklasse

Wort	abs. Häufigkeit	HK
der	7507542	0
die	6836196	0
und	4965269	1
in	3850950	1
den	2758375	1
von	2232692	2
zu	2081975	2
das	1889280	2
mit	1843993	2
sich	1751631	2
nicht	1680592	2
des	1625105	2
ist	1603975	2
auf	1595644	2
für	1584354	2
im	1535076	2
dem	1463017	2
ein	1323984	2
eine	1229665	3
als	1081635	3

Tabelle 4.1.: Die 20 häufigsten Wörter und ihre Häufigkeitsklasse (HK)

Die relative Häufigkeit eines Wortes berechnen wir aus dem Verhältnis des häufigsten Wortes (das Wort *der*) zum betrachteten Wort. Aus dieser relativen Häufigkeit bilden wir Häufigkeitsklasse, indem wir diesen Wert logarithmieren und auf die nächste ganze Zahl runden:

$$HK(a) \approx log_2\left(h(\text{„der“})/h(a)\right)$$

Nach dieser Formel teilen wir die Wörter in die Häufigkeitsklassen 0 bis 22 ein. In der Häufigkeitsklassen 22 sind alle Wörter, die in Texten bis jetzt nur einmal gelesen wurden. Allgemein enthält eine Klasse $\mathcal{H}$ alle diejenigen Wörter, die ca. $2^{\mathcal{H}}$ mal seltener als das häufigste Wort („*der*“) in den Texten vorgekommen sind. In Tabelle 4.1 sind die häufigsten Wörter der deutschen Sprache zusammen mit ihrer Häufigkeitsklasse aufgeführt.

Sachgebiete

Die Sachgebietsangaben wurden aus verschiedenen Quellen zusammengetragen und werden gegenwärtig genormt und in einer Hierarchie geordnet. Angezeigt wird die unterste Hierarchieebene des Sachgebietsbaumes, in die das Wort eingeordnet ist[2]. Um die Sachgebietsstruktur komfortabel editieren und graphisch darstellen zu können, benutzen wir den Editor *MindMap*. Zu diesem Zweck exportieren wir die Hierarchie aus der relationalen Datenbank in eine Textdatei, die dann in *MindMap* importiert werden kann. Einen Ausschnitt aus der Darstellung der Sachgebietshierarchie durch dieses Programm zeigt Abbildung 4.10. Nachdem die Hierarchie bearbeitet worden ist, kann in analoger Weise wieder in die Datenbank importiert werden.

Momentan überführen wir die Bezeichnungen der Sachgebiete in die der Schlagwortnormdatei [DB97] der Deutschen Bibliothek. Andere Sachgebietsangaben, z. B. von feineren Unterteilungen aus Fachsprachen, werden in diese Hierarchie mit eingearbeitet.

Beschreibung und Pragmatikangaben

Die Beschreibungen und Pragmatikangaben werden so angezeigt, wie sie in den entsprechenden Tabellen vorliegen.

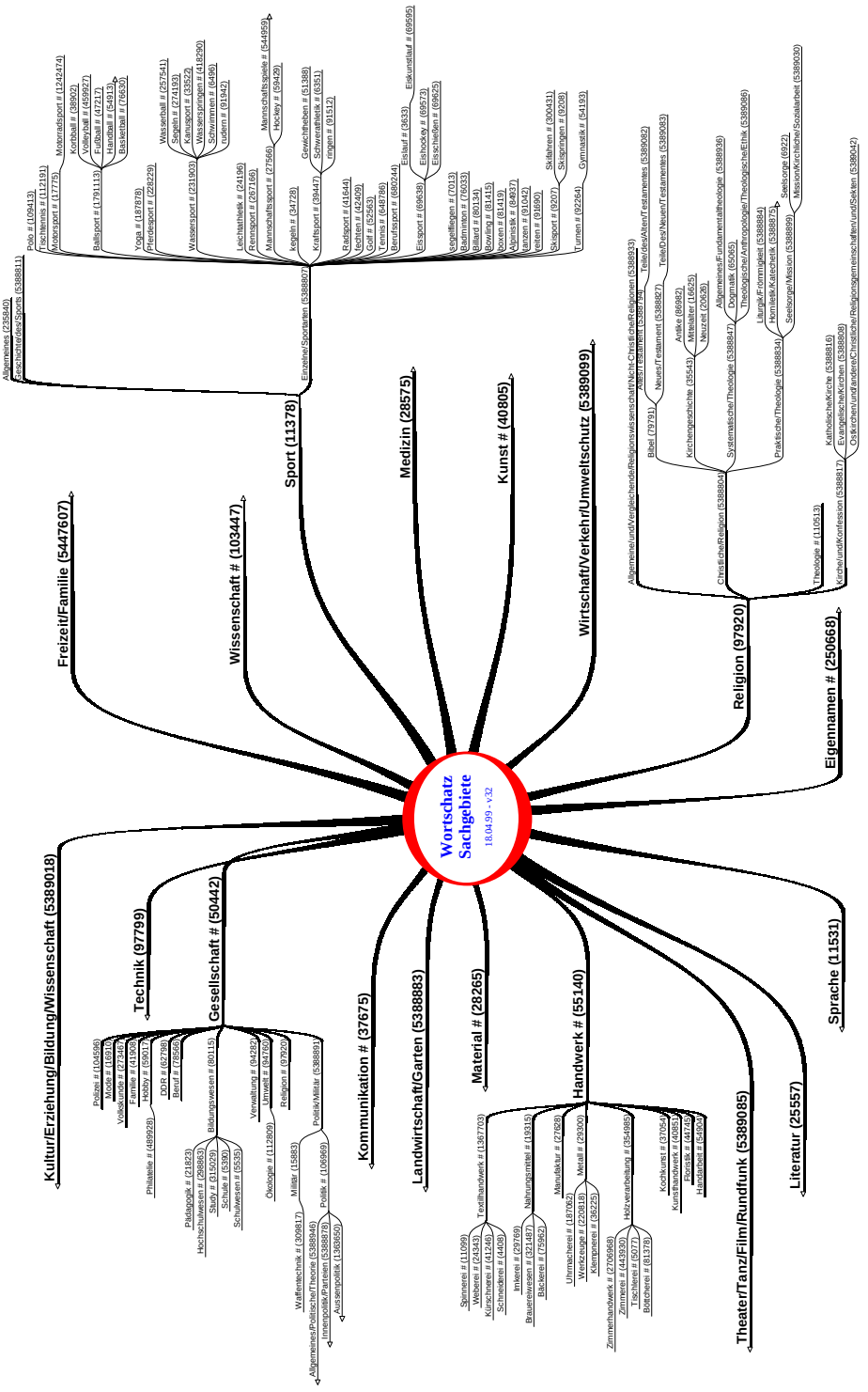
Es liegen 130.000 Beschreibungen vor, die aber auf Grund einer datenbanktechnischen Beschränkung aus den Anfangszeiten des Projekts auf eine Länge von 64 Zeichen beschränkt sind. Neue Beschreibungen unterliegen dieser Längenbeschränkung nicht mehr.

Zur Pragmatik liegen 34.000 Angaben vor, etwa `obersächs.` oder `gebr: derb abwertend`. 75% der Angaben aus verschiedenen Quellen sind bis jetzt auf eine einheitliche Bezeichnungsweise normiert.

Morphologie

Die morphologische Zerlegung der Wörter wurde mit dem eigens dazu entwickelten Programm *MorphDiv* durchgeführt. Das Programm basiert auf einer Liste zulässiger Morphempaarkombinationen. Diesen sind Informationen über den Typ der einzelnen Morpheme (Präfix, Stamm, Suffix, ...) sowie den Typ der angrenzenden Morpheme zugeordnet. Weiterhin besitzen die Morphempaare eine Übergangswahrscheinlichkeit, die aus Frequenzangaben von Trainingsdaten gebildet wird und für die Auswahl einer bestimmten Zerlegung herangezogen werden kann.

[2] natürlich kann ein Wort mehreren Sachgebieten angehören, so gehört *Valenz* den Sachgebieten *Physik, Chemie, Umwelt, Medizin* und *Theoretische und Physikalische Chemie* an



sachgebiete_da3.mmp - 18.04.99

Abbildung 4.10.: Ausschnitt aus der Sachgebietsstruktur

Tabelle 4.2 zeigt eine Auflistung der verwendeten Zeichen für die Morphemtypen.

Tabelle 4.2.: Morphemtypen

Kürzel	Erläuterung	Beispiele
Wortstamm		
=	normaler Stamm	Blut, priv, punkt, stimm, tens
(	französischer Stamm	ball, brill, cercl, drain, mass, moul, pass, tri
)	englischer Stamm	camp, check, coach, cup, spray, sex
Suffix		
%	normale Endung	a, e, en, em, er, es, et, ig, o, t, te, ten, ter, ung
~	französische Endung	ag, e, ee, ier, on
_	lateinische Endung	al, am, ant, at, gen, i, in, ior, it, iv, on, phob, ur
Präfix		
+	normales Präfix	a, ab, an, be, bi, dys, er, ex, kom, kon, re, tri
Abkürzung		
=^	normale Abkürzung (steht vor jedem Buchstaben der Abkürzung)	`=^b=^z=^w`

Eine detaillierte Beschreibung des Algorithmus findet sich in [Bö98].

Grammatikangaben

Die Grammatikangaben sind in der Datenbank kodiert angelegt, um Speicherplatz zu sparen. Da sie aber sehr vielfältig sein können, sind alle Angaben in einer Datenbankspalte als Zeichenkette gespeichert, für deren Format eine eigene Syntax entworfen wurde.

Den größten Umfang nehmen die Angaben zur Wortart und der Flexion der Wörter ein. Die gekennzeichneten Wortarten sind: Substantiv, Verb, Adjektiv, Adverb, Präposition, Interjektion und Pronomen, Außerdem wurden auf Grund des großen Bestandes Eigennamen gesondert markiert.

Angaben zum Flexionstyp liegen zu Substantiven und Adjektiven vor. Für Substantive werden bei der Anzeige des Wortes die gebeugten Formen aus der Stammform und einem Flexionsschema generiert, die in einer Datenbank-Tabelle zu den einzelnen Flexionstypen abgelegten Endungen an die Stammform des Substantives

angehängt werden.

Die verschiedenen Flexionstypen haben wir aus mehreren Quellen zusammengetragen und auf 76 Typen unifiziert. Die Angaben zur Flexion lassen sich auf zwei Arten automatisch bestimmen. Zum einen ist es durch die Sammlung von Vollformen möglich, die verwendeten Endungen eines Substantives zu bestimmen. Dadurch kann die Zahl der möglichen Typen meist auf wenige eingeschränkt werden, sofern das Wort und seine Beugungen oft genug im Korpus auftauchen.

Zum anderen können die Angaben eines Wortes auf ein anderes übertragen werden, wenn eine gewisse Anzahl der letzten Buchstaben beider Wörter übereinstimmen. Das ist möglich, weil Wörter mit der gleichen Endung (wie *-schaft, -heit, -tion*) gleich flektiert werden. Außerdem übernimmt ein zusammengesetztes Wort die grammatikalischen Eigenschaften des Grundwortes, d. h. ihres letzten Bestandteils. Bei der Übernahme von Flexion und Geschlecht stellte sich heraus, dass bei der Übereinstimmung der letzten fünf Buchstaben noch eine zuverlässige Übernahme garantiert werden kann.

Da wir über Flexionsangaben aus verschiedenen Quellen verfügen, werden diese vor der Anzeige gegeneinander überprüft. Wenn zu einem Wort verschiedene Flexionsarten bestehen (z. B. bei Homonymen: *Bank – Banken, Bank – Bänke*), werden diese bei der Anzeige ergänzt. Falls in den Grammatikangaben verschiedene Varianten angegeben werden, weil zum Zeitpunkt der Erzeugung der Flexionstyp nicht eindeutig bestimmt werden konnte, wird aus diesen Varianten die Schnittmenge gebildet.

Beispiel:
Zu einem Wort existieren die Flexionsangaben „*a1 oder a3; c2*“, „*a1 oder a2*“ und „*c4*“. Aus den ersten beiden Angaben können wir schließen, dass das Wort nach Typ *a1* flektiert wird. Nach der ersten und dritten Angabe kann es weiterhin nach *c2* oder *c4* flektiert werden. Also werden die Formen nach den Typen *a1, c2* und *c4* generiert.

Neben Angaben zur Wortart und Flexion liegen noch weitere Angaben vor:

- für Substantive: Genus, Numerusgebrauch (z. B. *singulare tantum*)
- für Verben: Stammform, Partizipbildung (mit *haben* oder *sein*), transitiv/intransitiv, abtrennbares Präfix, reflexiv, und Verbrahmen (zum automatischen Parsen von Sätzen)

Relationen zu anderen Wörtern

Aus externen Quellen wurden in das Worschatz-Lexikon Relationen zwischen Wörtern übernommen. Dazu zählen Synonyme, Antonyme und nicht näher spezifizierte Relationen (in Wörterbüchern z. B. mit *siehe auch* bezeichnet).

Oft liegen Relationen nicht nur zwischen zwei, sondern mehreren Wörtern vor, z. B. sind Synonyme von *Quark: Quatsch, Schichtkäse, Topfen* und *Weißkäse*. Zwar sind *Schichtkäse* und *Weißkäse* Synonyme, dies gilt jedoch nicht für *Quatsch* und *Schichtkäse*. Dies resultiert aus der Tatsache, dass *Quark* verschiedenen Homonymgruppen angehört. Deshalb wird zu solchen Relationsgruppen ein Wort gesondert als *Kopf* der Gruppe gekennzeichnet gespeichert. Zum Kopf einer Relationsgruppe werden alle anderen Mitglieder dieser Gruppe angezeigt („*Synonyme: ...* "), zu einem Gruppenmitglied jedoch nur ausgegeben, dass eine umgekehrte Referenz besteht („*ist Synonym von: ...* ").

Die Wörter aus den Relationen werden mit Hyperlinks verbunden, die auf die Angaben zu dem Wort in dieser Schreibweise verweisen, denn den Nutzer interessiert sich, wenn er von *Quark* zu *Topfen* klickt, nicht für das Verb *topfen* (*Pflanzen umtopfen*).

Weitere Angaben zum Wort

Zu jedem Wort werden drei Beispielsätze mit Quellenangabe angezeigt. Über einen Link sind für autorisierte Nutzer jeweils zehn weitere Sätze erreichbar. Auf das Konzept der Nutzerberechtigung wird im Abschnitt Sicherheitskonzept (s. u.) eingegangen.

Weiterhin hat der Nutzer die Möglicheit, sich die Kollokationen nach den Maßen $\sigma_{CBA}, \sigma_{CBA\_nbli}$ und σ_{CBA_nbre} mit den zugehörigen Signifikanzwerten auf einer gesonderten anzeigen zu lassen (siehe dazu Abschnitt 3.1.2 auf Seite 25). Darüber hinaus wird auch der Kollokationsgraph (siehe Abschnitt 4.3 auf Seite 53) angezeigt und die Kollokationen zweiter Ordnung, sofern sie schon berechnet wurden.

Auf der WWW-Seite, die die mit den Informationen zum Wort enthält, wählt die Nutzerin, welche Informationen zu den Kollokationen sie angezeigt bekommen möchte. Zusammengefasst sind Kollokationen erster Ordnung, der Kollokationsgraph und die Kollokationen zweiter Ordnung. Diese Auswahl gilt die Anzeige aller weiteren Kollokationen, die über Links der ersten Kollokationen ausgewählt wurden.

Um einen wiederholten Zugriff auf den Kollokationsgraphen zu beschleunigen, wird er eine gewisse Zeit auf dem Server zwischengespeichert, anderenfalls wird er neu generiert. Für die Kollokationen zweiter Ordnung ist eine Generierung auf Anforderung geplant. Eine Vorausberechnung ist hier nicht sinnvoll, da die Speicherung aller Kollokationen zweiter Ordnung sehr speicheraufwendig ist.

Dynamische Anfragetypen

Anfragen, die sich auf eine Tabelle beschränken, können der WWW-Oberfläche des Wortschatz-Projektes leicht hinzugefügt werden. Dazu wird die zugehörige SQL-

Anfrage mit einer Überschrift in einer gesonderten Datenbank-Tabelle (`abfragen`) abgelegt. In der SQL-Anfrage sind die variablen Stellen durch geschweifte Klammern gekennzeichnet. Die Nummer der Abfragen dient außerdem dazu, die Anfragen für statistische Zwecke mit zu protokollieren.

Die Überschriften werden automatisch in einem Auswahlmenü angeboten. Wählt man eine Überschrift, wird eine Anfrageseite automatisch erzeugt. In dieser Seite werden Eingabefelder für die variablen Felder angezeigt. Die Beschriftung der Eingabefelder ist der Name, der in der Tabelle der Anfragen in geschweiften Klammern steht. Er wird auch als Beschriftung der Anfrageseite und als Variablenname des CGI-Programms verwendet. Die Anfrageseite kann als Rohgerüst genutzt werden, um in einer Kopie der Seite die Anfrage zu kommentieren oder grafisch ansprechender zu gestalten.

In der Ergebnis-Seite werden die ersten 20 Zeilen der Datenbank-Rückgabe dargestellt. Wenn Datenbankspalten Wortnummern beinhalten, werden diese nicht dargestellt, sondern die folgende Spalte mit Links auf das zugehörende Wort versehen. Über zwei weitere Links kann auf die 20 vorhergehenden oder folgenden Zeilen zugegriffen werden.

Testumgebung für Kollokationen zweiter Ordung

Der Link *Relationen der Testwörter* dient für eine Testumgebung der Kollokationen zweiter Ordnung (siehe Abbildung 4.11). Im oberen Bereich kann man ein Wort aus den Bereichen *Stoppwörter, Mehrwortphrasen, Medizin, Informatik, Politik, Recht, Sport* und *Allgemeiner Sprachgebrauch* auswählen. Für diese 78 Wörter haben wir alle Kollokationen zweiter Ordnung berechnet, um die Signifikanzmaße bewerten zu können und um Kombinationen der Maße zu probieren.

Die untere Bereich der Eingabeseite dient zur Auswahl der anzuzeigenden Kollokationsmaße, zur Angabe von Kombinationen dieser Maße, von Zusatzbedingungen und Ordnungskriterien. Die Zusatzbedingungen können auch auf andere Tabellen der Datenbank zugreifen.

Sicherheitskonzept

Die WWW-Oberfläche des Wortschatz-Projektes wird sowohl von externen Nutzern als auch den Projektmitarbeitern genutzt. Diesen stehen auch Informationen zur Verfügung, die noch nicht für externe Nutzer aufbereitet sind. Außerdem sind eingeschränkte Editiermöglichkeiten implementiert, die wir weiter ausgebauen wollen. Außerdem sollten Nutzer Zugriff auf noch nicht öffentliche Daten erhalten, diese aber nicht ändern können.

Abbildung 4.11.: Testumgebung für Kollokationen zweiter Ordung

Aus diesen Gründen wurde in die Wortschatz-Oberfläche ein mehrstufiges Sicherheitskonzept eingebaut. Dazu werden im Quellcode mit Compilerdefinitionen Programmteile je nach Nutzergruppe aus- oder eingeblendet, um bestimmte Funktionen nur in einer Instanz des Programmes freizuschalten oder je nach Nutzergruppe verschiedenen Programmcode auszuwählen.

Diese Programme werden in verschiedenen Pfaden des WWW-Servers abgelegt. Für diese Pfade sind im WWW-Server unterschiedliche Rechte eingerichtet. Je nach Berechtigungsebene ist eine Authentifizierung des Nutzers notwendig und/oder der Zugriff auf einige ausgewählte Rechner beschränkt.

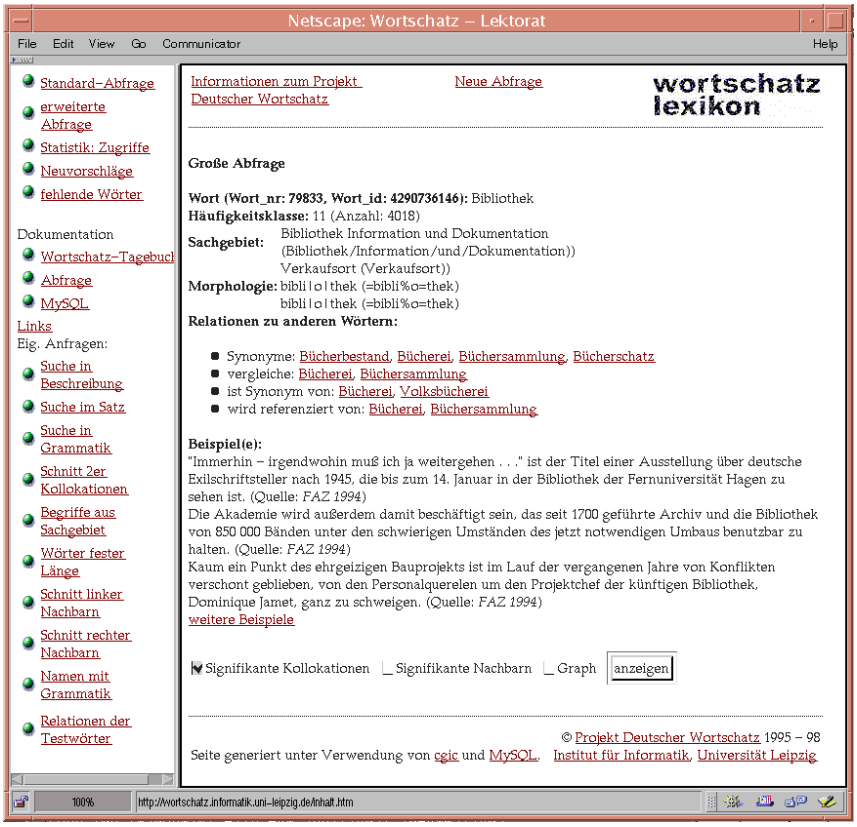


Abbildung 4.12.: Ergebnis der erweiterte Abfrage für *Bibliothek*

Eine weitere Sicherheitsfunktion dient als Schutz gegen ein automatisches Herunterladen der Wortschatz-Daten. Dazu wird die Anzahl der Zugriffe je Nutzer je Rechner, von dem der Zugriff erfolgte, in der Datenbanktabelle `ip_log` gezählt. Wenn aus dem Teilnetz, dem der Nutzers angehört, innerhalb der letzten 24 Stunden eine bestimmte Anzahl an Zugriffen (z. Zt. 250) überschritten wurde, wird dieser Rechner in der Tabelle `enemy` vermerkt und jede Abfrage um 60 Sekunden verzögert. Wenn aus dem Teilnetz einen Tag lang nicht zugegriffen wurde, erlischt die Verzögerung. Die Zahlenwerte geben lediglich die momentane Konfiguration an.

Weiterhin wird gespeichert, zu welchen Wörtern Anfragen gestellt wurden. Dadurch kann das Verhalten des Abfrageprogramms auf die Vorstellungen der Nutzer ausgerichtet werden. Zu diesen Einstellungen zählen z. B. die Wahl der Jokerzeichen, Eingabe der Umlaute von Rechnern ohne deutsche Tastatur, Abfrage von mehreren Wörtern (nicht implementiert) und Einbau einer Rechtschreibkorrektur.

5. Zusammenfassung

Ziel der vorliegenden Arbeit war die automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten. Zunächst lag der Schwerpunkt dabei auf der Extraktion von Kollokationen aus dem Textkorpus. Hierfür wurden verschiedene aus der Literatur bekannte Verfahren auf ihre Anwendbarkeit hin geprüft und ihre Leistungsfähigkeit verglichen. Es stellte sich heraus, dass keines der getesteten Verfahren den gestellten Ansprüchen genügte. Dies motivierte uns zur Entwicklung eines eigenen Verfahrens, welches auf dem Common-Birthday-Problem aus der mathematischen Statistik basiert.

Dieses neue Signifikanzmaß lieferte zuverlässig gute Werte zur Berechnung der Kollokationen und verfügt über eine Reihe vorteilhafter Eigenschaften, die in Kapitel 3 ausführlich dargestellt wurden. Eine bereits im Rahmen des Wortschatz-Projekts entwickelte, effiziente Implementierung eines Suchbaums für die Auffindung von Wortpaaren und deren Häufigkeiten ermöglichte es, die Kollokationen für alle Wörter des Textkorpus zu berechnen. Infolge dessen sind jetzt zu den meisten Wörtern Kollokationsangaben verfügbar, was die Attraktivität der Datenbank steigert. Ausgenommen bei der Berechnung wurden lediglich Stoppwörter und sehr seltene Wörter, da bei denen a priori klar war, das keine sinnvollen Kollokationen existieren können.

Aufbauend auf die Kollokationsuntersuchungen ergaben sich neue Ansätze zur Berechnung von semantischen Relationen, die interessante Ergebnisse lieferten und eine weitere Forschung in diesem Gebiet vielversprechend erscheinen lassen.

Ein weiterer Schwerpunkt der Arbeit lag in der Visualisierung der berechneten Daten, um z. B. deren Qualität einschätzen zu können. Da dies mit mit einer textbasierten Darstellung nur eingeschränkt möglich ist, wurde eine graphische Ausgabe angestrebt, die sowohl ausdruckskräftig ist, als auch ästhetischen Ansprüchen genügt. Mit dem implementierten Simulated-Annealing-Algorithmus konnten bei geringem Ressourcenverbrauch mit akzeptabler Geschwindigkeit Graphen erzeugt werden, die diese Vorgaben erfüllen. Eine weitere Optimierung der Darstellung wäre nach Davidson & Harel auf Kosten der Berechnungseffizienz möglich.

Die Verfahren wurden im Rahmen des Wortschatz-Projektes implementiert und stehen sowohl Projektmitarbeitern am Institut als auch interessierten externen Nutzern

zur Verfügung. Seit der Verfügbarkeit der Implementierung werden die Werkzeuge und die von ihnen generierten Daten regelmäßig verwendet und haben damit ihre Praxistauglichkeit unter Beweis gestellt.

Literaturverzeichnis

[CT94] Isabel F. Cruz, Roberto Tamassia: *How to Visualize a Graph: Specification and Algorithms. Part I: Algorithmic Approach*, http://www.cs.brown.edu/people/rt/gd-tutorial.html, 1994

[Bö98] Timo Böhme: *Morphologische Zerlegung. Dokumentation zu MorphDiv*, Universität Leipzig, 1998

[DH96] Ron Davidson, David Harel: *Drawing Graphs Nicely Using Simulated Annealing*, in: ACM Transactions on Graphics, Vol. 15, No. 4, S. 301-331, 1996

[Do64] Franz Dornseiff: *Sprache und Sprechender*, Leipzig, 1964

[DB97] *Normdaten-CD-ROM: Gemeinsame Körperschaftsdatei, Personennamendatei, Schlagwortnormdatei.* Die Deutsche Bibliothek, Frankfurt am Main, 1997

[DuInf88] *Duden «Informatik»: ein Sachlexikon für Studium und Praxis.* hrsg. vom Lektorat d. BI-Wiss.-Verl. unter Leitung von Hermann Engesser. Bearb. von Volker Claus u. Andreas Schwill; Mannheim, Wien, Zürich, 1988

[Ea84] Peter Eades: *A Heuristic for Graph Drawing*, Congressus Numerantium, Bd. 42, S. 149-160, 1984

[Hm85] Franz-Josef Hausmann: *Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels*, in: Henning Bergenholtz, Joachim Mugdan (Hrsg.): *Lexikographie und Grammatik: Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 1984 (Lexicographica 3)*, Niemeyer; Tübingen, S. 118-129, 1985

[La96] Stefan Langer: *Selektionsklassen und Hyponymie im Lexikon*, Universität München, Centrum für Informations- und Sprachverarbeitung; München, 1996

[Lr96] Lehr, Andrea: *Kollokationen und maschinenlesbare Korpora : ein operationales Analysemodell zum Aufbau lexikalischer Netze*, Niemeyer; Tübingen, 1996.

[Lm97] Lothar Lemnitzer: *Komplexe lexikalische Einheiten in Text und Lexikon*, in: Gerhard Heyer, Christian Wolff (Hrsg.): *Linguistik und neue Medien. Tagungsband der 10. Jahrestagung der Gesellschaft für linguistische Datenverarbeitung*, Universität Leipzig, 1998

[Ml76] Igor A. Mel'čuk: *Towards a linguistic „Meaning-text" model*, in: *Das Wort*, S. 26-62. Fink; München, 1976

[Ra97] Friedhelm Ramme: *Transparente und effiziente Nutzung partitionierbarer Parallelrechner*, Logos Verlag; Berlin, 1997

[Rp96] Reinhard Rapp: *Die Berechnung von Assoziationen: ein korpuslinguistischer Ansatz.* Olms; Hildesheim, Zürich, New York, 1996 (`http://www.fask.uni-mainz.de/user/rapp/papers/disshtml/main/node3.html`)

[Ru95] Gerda Ruge: *Wortbedeutung und Termassoziation: Methoden zur automatischen semantischen Klassifikation*, Olms; Hildesheim, New York, Zürich, 1995

[SG83] Gerard Salton, Michael J. McGill: *Information Retrieval: Grundlegendes für Informationswissenschaftler*, McGraw-Hill, Hamburg, 1983

[Sm86] Helmut Schumacher (Hrsg.): *Verben in Feldern: Valenzwörterbuch zur Syntax und Semantik deutscher Verben*, de Gruyter; Berlin, New York, 1986

[St90] James Steele: *Meaning-text theory: linguistics, lexicography and implications*, University of Ottawa Press; Ottawa, London, Paris, 1990

[Ti99] Lydia Thießen: *Substantiv-Adjektiv-Kollokationen*, Universität Leipzig, 1999

[WW98] Elke Warmuth, Walter Warmuth *Elementare Wahrscheinlichkeitsrechung: vom Umgang mit dem Zufall*, Teubner; Stuttgart, Leipzig, 1998

[Wt85] Hermann Witting: *Mathematische Statistik, Band 1: Parametrische Verfahren bei festem Stichprobenumfang* Teubner; Stuttgart, 1985

A. Lexikalische Funktionen

A.1. Paradigmatische Funktionen

A.1.1. Substitutionen

Funkt.	Beschreibung	Beispiel
Logische Grund-Substitutionen		
Syn	Synonyme	Syn(Piano) = Klavier
Anti	Antonyme	Anti(präventiv) = postmortal
Conv	Konverse (Permutation der Argumente)	Conv213(geben) = nehmen, Conv213(verkaufen) = kaufen
Da es kaum „reine“ Synonyme gibt, werden zu diesen Relationen zusätzliche Informationen gespeichert, inwieweit die Lexeme hinsichtlich ihrer Bedeutung verändert sind oder beispielsweise nur regional oder in bestimmten Kulturkreisen üblich sind		
Wortableitungen		
S0	Ableitung eines Substantivs	S0(drücken) = Druck
V0	Ableitung eines Verbs	V0(Druck) = drücken
A0	Ableitung eines Adjektives	A0(glätten) = glatt
Adv0	Ableitung eines Adverbs	Adv0(sichern) = sicher
Diese Ableitungen werden unterteilt in morphologische (schön → Schönheit) semantische Wortableitungen (Frankreich → französisch)		
Kontrast-Terme		
Contr	Kontrast	Contr(schwarz) = weiß, Contr(rechts) = links

A.1.2. Qualifier

Funkt.	Beschreibung	Beispiel
Bewertungen		
Magn	Größe	Magn(Temperatur) = heiß

Funkt.	Beschreibung	Beispiel
Bon	Bonität	Bon(Argument) = stark
Ver	default-Wert	Ver(Messer) = scharf
Generische Kategorien		
Gener	generische Kategorie, Obergegriff	Gener(Schmerz) = Gefühl
Mengenbeziehungen		
Mult	Zusammenfassungen	Mult(Haare) = Büschel
Sing	Elemente	Sing(Regen) = Tropfen
Organisations-Beziehungen		
Cap	Leiter	Cap(Schiff) = Kapitän, Cap(Stadt) = Bürgermeister
Equip	Mitarbeiter	Equip(Verein) = Mitglieder
Größenänderung		
Nur in Kombination mit anderen Funktionen, speziell Pred, gebraucht		
Plus	Vergrößerung	PredPlus(Aufmerksamkeit) = erhöhen
Minus	Verminderung	PredMinus(Interesse) = nachlassen
Übertragene Bedeutung		
Figur	Standard-Metapher, in Kombination mit dem Argument ergibt sich ein eingeschränktes Synonym	Figur(Verzweiflung) = tiefes Gefühl (der Verzweiflung), Figur(Tag) = (Tages-)Licht
Pleonastische Adjektive		
Epit	Standard-Adjektiv, dessen Bedeutung bereits im Argument enthalten ist (Epitheton)	Epit(Pfarrer) = geistlich, Epit(Schimmel) = weiß

A.1.3. Aspekte der Situation

Funkt.	Beschreibung	Beispiel
Prozeß		
Caus	Grund	CausFunc0(Hoffnung) = wecken
Germ	Keim, Beginn	Germ(Bach) = entspringen
Culm	Höhepunkt	Culm(Turnier) = gewinnen
Degrad	Verschlechterung	Degrad(Farbe) = ausbleichen
Excess	Übermaß	Excess(Auto) = rennfahren
Obstr	mit Schwierigkeiten	Obstr(Redner) = stottern
Prejor	an Wert verlieren	Prejor(Aktien) = fallen
Liqu	Liquidieren	Liqu(Schmerz) = überwinden

Funkt.	Beschreibung	Beispiel
Perf	Zustand am natürlichen Ende des Prozesses	S1Perf(sterben) = Verstorbener
Result	Ergebnis des Prozesses	Result(Aufstehen) = stehen
Phasen		
Incep	Beginn	IncepPred(krank) = erkranken
Cont	Verlauf	ContFunc0(Angebot) = aufrechterhalten
Fin	Ende	FinOper1(Gedächtnis) = verlieren
Teilnahme		
Involv	Verb, verknüpft mit dem Argument, Nebenhandlung	Involv(Ton) = (den Raum) füllen
Instr	Präposition, mit der das Argument als Instrument benutzt wird	Instr(Fuß) = zu (Fuß)
Manif	Sichtbar werden, oft zusammen mit Caus	Caus1Manif(Meinung) = ausdrücken
Nocer	schädlich sein für	Nocer(Angst) = lähmen
Perm	Erlauben	nonPerm1Manif(Gefühl) = verstecken
Prepar	Vorbereiten	PreparOper1(Gewehr) = laden
Propt	Präposition, mit der das Argument als Grund genutzt wird	Propt(Ehrfurcht) = aus
Prox	zeitliches oder räumliches Angrenzen	Prox(abfliegen) = sich am Startplatz befinden
Sympt	Symptom sein für etwas	Sympt(Neid) = grün anlaufen (vor Neid)
Son	typischen Sound erzeugen	Son(Hund) = bellen
S-instr	Standard-Name für Instrument	S-instr(schneiden) = Messer
S-med	Standard-Name für Medium	S-med(sprechen) = Stimme
S-mod	Standard-Name für Modus	S-mod(bezahlen) = bar, per Scheck, ...
S-loc	Standard-Name für Ort	S-loc(Häftling) = Gefängnis
S-res	Standard-Name für Resultat	S-res(Bauer) = Ernte
Raum-zeitliche Eigenschaften		
Loc-in	Präposition, befindet sich in	Loc-in(Eintfernung) = in (einer Entfernung)
Loc-ab	Präposition, Fortbewegung von	Loc-ab(Entfernung) = aus (einer Entfernung)
Loc-ad	Präposition, Bewegung hin zu	Loc-ad(Platz) = zu (einem Platz)

Funkt.	Beschreibung	Beispiel
Loc-*-temp	analog mit zeitlicher Bedeutung	Loc-in-temp(Morgen) = am (Morgen)
Centr	in der Mitte	Centr(Rad) = Achse
Standard-Namen für Teilnehmer		
S1, S2, ...	für die entsprechend numerierten Teilnehmer einer Aktion	S1(unterrichten) = Lehrer, S2(unterrichten) = Schüler
Standard-Kommandos für Teilnehmer		
Imper	Kommando	Imper(Schweigen) = Ruhe!
Kopula		
Copul	Koplula	Copul(Warnung) = dienen als (Warnung)

A.1.4. Qualifier für Aktanten

Funkt.	Beschreibung	Beispiel
Typische Qualifier für Aktanten		
A1, A2, ...	Adjektive für die entsprechend numerierten Aktanten einer Aktion	A1(Liebe) = verliebt, A2(Liebe) = geliebt
Adv1, ...	Adverbien für die entsprechend numerierten Aktanten einer Aktion	
Spezielle Qualifier für Aktanten		
Able1, ...	Adjektive für die entsprechend numerierten Aktanten einer Aktion, die spezielle Fähigkeiten ausdrücken	Able1(lesen) = des Lesens kundig, Able2(lesen) = lesbar
Qual1, ...	Adjektive für die entsprechend numerierten Aktanten, die spezielle Eigenschaften für den Erfolg ausdrücken	Qual2(glauben) = plausibel
Pos1, ...	positive Adjektive für die entsprechend numerierten Teilnehmer	Pos2(Eindruck) = gut

A.2. Syntagmatische Funktionen

A.2.1. Verbale Operatoren

Funkt.	Beschreibung	Beispiel
Semantisch leere verbale Operatoren		
Oper1, ...	Verb, welches den entsprechenden Aktanten als sein grammatisches Subjekt nimmt und das Argument als (Akkusativ-)Objekt	Oper1(Angebot) = (ein Angebot) machen, Oper2(Angebot) = (ein Angebot) bilden, ausmachen
Func1, ...	Verb, welches den entsprechenden Aktanten als sein grammatisches (Akkusativ-)Objekt nimmt und das Argument als Subjekt	Func0(Sturm) = sein, Func1(Angebot) = kommen (von jemandem), Func2(Angebot) = betreffen (etwas)
Laborij	Verb, welches die entsprechenden Aktanten i und j als sein grammatisches Subjekt und (Akkusativ-) Objekt nimmt und das Argument als (Dativ-)Objekt	Labor12(Risiko) = aussetzen (jemanden einem Risiko)
Semantische verbale Operatoren		
Real1, ...	Verb, welches den entsprechenden Aktanten als sein grammatisches Subjekt nimmt und das Argument als (Akkusativ-)Objekt	Real1(Problem) = lösen, Real2(Prüfung) = bestehen
Fact1, ...	Verb, welches den entsprechenden Aktanten als sein grammatisches (Akkusativ-) Objekt nimmt und das Argument als Subjekt	Fact0(Flugzeug) = fliegen
Labrealij	Verb, welches die entsprechenden Aktanten i und j als sein grammatisches Subjekt und (Akkusativ-) Objekt nimmt und das Argument als (Dativ-)Objekt	Labreal12(Reservierung) = halten (etwas in Reserve)

A.2.2. Prädikatoren

Funkt.	Beschreibung	Beispiel
Pred	Verbalisierung von Nomen oder Adjektiven	Pred(Abstinenzler) = sich enthalten, CausePred(dunkel) = verdunkeln, IncepPred(dunkel) = dunkeln, IncepPredMinus(Schmerz) = nachlassen

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Leipzig, am 19. 4. 1999

Fabian Schmidt